Exploiting Social and Topical Context for Predicting User Preference in Microblogging

Ye Wu

A Dissertation Presented in Partial Fulfillment of the
Requirement for the Degree Doctor of Philosophy

2014

The University of Tokushima
Graduate School of Engineering
Information Science and Systems Engineering

Contents

1 Introduction **1**
- 1.1 Background . . . 2
  - 1.1.1 User Preference in Microblogging . . . 2
  - 1.1.2 User Interest Prediction . . . 3
  - 1.1.3 User Opinion Prediction . . . 4
  - 1.1.4 Exploiting Social and Topical Context . . . 5
- 1.2 Contributions . . . 8
- 1.3 Organization . . . 8

2 Literature Review **10**
- 2.1 Recommender Systems . . . 10
  - 2.1.1 Techniques of General Recommender Systems . . . 11
  - 2.1.2 Recommender Systems Using Context Information . . . 13
  - 2.1.3 Personalized Recommendation in Microblogging . . . 15
- 2.2 Research about Opinion Prediction . . . 16
  - 2.2.1 Sentiment Analysis and Opinion Mining . . . 16
  - 2.2.2 Sentiment Analysis in Microblogging . . . 17
  - 2.2.3 Detecting Opinion from Microblogging for Applications . . . 19

3 Data Collection and Preprocessing **21**
- 3.1 Data Collection . . . 21
  - 3.1.1 Why Twitter . . . 21
  - 3.1.2 Crawling Data with Twitter API . . . 23
- 3.2 Data Preprocessing . . . 24
  - 3.2.1 Language Selection . . . 24
  - 3.2.2 User Filtering . . . 24
  - 3.2.3 User-topic Opinion Labeling . . . 24

4 Social and Topical context incorporated Framework **27**
- 4.1 The Basic Low-rank Matrix Factorization Model . . . 27
- 4.2 Social Context Regularization . . . 29
  - 4.2.1 Social Context Hypothesis . . . 30
  - 4.2.2 Exploiting Social Context for Regularization . . . 31
- 4.3 Topical Context Regularization . . . 33
  - 4.3.1 Topical Context Hypothesis . . . 34
  - 4.3.2 Exploiting Topical Context for Regularization . . . 34
- 4.4 ScTcMF: The Proposed Framework with Social and Topical Context 37

5 User Interest Prediction **39**
5.1 Problem Definition . . . 39
5.2 Exploiting Social and Topical Context for Predicting User Interest . . 40
5.2.1 The Selected Dataset for User Interest Prediction . . . 40
5.2.2 Data Observations on Topical Opinion Distribution . . . 41
5.2.3 Data Analysis on User Interests . . . 44
5.2.4 Incorporating Social Context . . . 46
5.2.5 Incorporating Topical Context . . . 48
5.2.6 Details of ScTcMF Algorithm Solution . . . 49
5.3 Experiments on User Interest Prediction . . . 52
5.3.1 Experiment Setup . . . 52
5.3.2 Performance Comparison of User Interest Prediction . . . 53
5.3.3 Time Complexity and Runtime Convergence . . . 56
5.3.4 Effects of Social and Topical Context Regularization . . . 58

6 User Opinion Prediction **61**
6.1 Problem Definition . . . 61
6.2 Exploiting Social and Topical Context for Predicting User Opinion . . 62
6.2.1 Incorporating Social Context . . . 63
6.2.2 Incorporating Topical Context . . . 64
6.2.3 Details of ScTcMF Algorithm Solution . . . 65
6.3 Experiments on User Opinion Prediction . . . 65
6.3.1 Experiment Setup . . . 66
6.3.2 Hypotheses Testing . . . 68
6.3.3 Performance Comparison of User Opinion Prediction . . . 69
6.3.4 Analysis and Discussion on User Opinion Prediction . . . 73
6.3.5 Parameter Analysis . . . 77

7 Conclusion and Future Work **82**
7.1 Conclusion . . . 83
7.2 Future Work . . . 83

List of Figures

1.1 The two tasks for predicting user preference. 3
1.2 A toy example of user preference prediction. 6

3.1 An example of Twitter homepage. 22
3.2 A quick test example of SentiStrength. 26

4.1 The process representation of basic low-rank matrix factorization. . . 29
4.2 *Social friend* relationship network in social context hypothesis. 30
4.3 The process representation exploiting social context. 32
4.4 The process representation exploiting topical context. 35

5.1 Popularity vs. Opinion Entropy 42
5.2 Ave Negative Strength vs. Ave Positive Strength 43
5.3 Average Absolute Opinion Strength vs. Subjective Ratio 44
5.4 Precision performance of user interest prediction. 55
5.5 Recall performance of user interest prediction. 56
5.6 Runtime convergence of the ScTcMF method. 57

6.1 User-topic level opinion prediction. 62
6.2 Accuracy comparisons using different training sets. 72
6.3 Impact of parameters α and β on RMSE. 78
6.4 Impact of parameters α and β on Accuracy. 79

List of Tables

2.1 List of emoticons verified by Twitter API. 18

3.1 The profiles of Twitter Dataset. 23
3.2 The list of main lexicons and linguistic rules. 25

5.1 Statistics of the dataset for user interest prediction. 40
5.2 Topic patterns based on opinion distribution. (L=Low, H=High.) . . 45
5.3 Statistics of user interest distribution difference. 46
5.4 The average means and variances of user interest similarities. 47
5.5 The statistical effects of social context regularization. 58
5.6 The statistical effects of topical context regularization. 59
5.7 The statistical results of ScTcMF vs. NMF 59

6.1 Statistics of the dataset for user opinion prediction. 67
6.2 RMSE comparisons using different training sets. 71
6.3 Precision comparisons in positive and negative opinion prediction . . 74
6.4 Recall comparisons in positive and negative opinion prediction 75
6.5 F1-Score comparisons in positive and negative opinion prediction . . 76

Acknowledgment

First and foremost, I would like to thank my advisor, Prof. Fuji Ren, who has been a invaluable mentor guiding me in my research. With his consistent support, I learned to open up new perspectives to address those challenging problems. More importantly, I learned from him the work attitude, ethic and disciplines that will benefit all my life.

I would like to thank my thesis committee, Dr. Kenji Terada and Dr. Masami Shishibori, for the assistance and encouragement that they have provided at all levels of my research. I also would like to thank the professors in Department of Information Science and Intelligent Systems, and the staffs in Center for International Cooperation in Engineering Education, who are very kind and patient to international students. They gave me a lot of help when I participated in the Double-Degree program in Tokushima.

I would like to thank the members of A1 group in The University of Tokushima. It is a great pleasure working with them, especially, Haitao Yu, Xin Kang, Song Liu, Jun Wang, Changqin Quan, Yan Sun, and Ji Li. It would be tough and harsh without their helpful suggestions and constant support.

Last but not least, I would like to thank my parents who have provided me with their love and affection and have believed in my abilities. They have been a pillar of strength behind me through the years, allowing me to focus and achieve my goals. In particular, I acknowledge my husband, Huiji Gao, without whose love and encouragement, I would not have finished this dissertation.

Abstract

Nowadays, microblogging service provides the rapidly updated information and online trends, which enriches and benefits people's daily life. Every day, hundreds of millions of people post their statuses and share information with 140-character limit short messages on the most popular microblogging services Twitter. On microblogging, users share and access fresh information in a more simple and convenient way, making the large amount of user generated data available.

However, the increasing topics of the posted messages also bring out an overload problem of information. To find out the really interesting topics for online users, not only helps users get out of this information overload trouble, but also improve the user experience of microblogging service. Therefore, the user interest prediction task is proposed to solve the problem "which topics are interesting to user".

Moreover, the user generated data in microblogging is also a resource including peoples' opinion information. How to infer microblogging users' opinions toward those topics they are interested in, in order to understand users further, is a very challenging problem but in demand in application scenarios such as viral marketing, opinion polling, mood monitoring, and so on. In the user opinion prediction task, we attempt to solve "what opinion does user hold on a specific topic".

In this dissertation, both the user interest prediction task and the user opinion prediction task are referred as user preference prediction, for user interest and user opinion represent user preference from different aspects. We focus on exploiting social and topical context to provide solutions for the two user preference prediction tasks. After capturing social and topical context information from microblogging data, we formulate it into the basic low-rank matrix factorization model, and finally propose the Social context and Topical context incorporated Matrix Factorization (ScTcMF) framework. The experimental results on the collected real-world Twitter

dataset demonstrate that social and topical context can lead to improvements in the performance evaluation, and the proposed ScTcMF framework can outperform the state-of-the-art methods in both user interest prediction and user opinion prediction.

Chapter 1

Introduction

In the age of Web 2.0, microblogging has become very popular and changed the way people interact with each other. In the fast-paced daily life, microblogging services allow users to share and receive information simply and efficiently. Users are able to create short messages on their home timeline. The topics of the posted messages range from the simple, such as "what I'm doing right now", to the thematic, such as "iphone", "world cup". Besides, commercial posts also exist for promotion and branding in microblogging websites. The rapid updated streams of microblogging posts provide more powerful and convenient access to information for people.

However, There is an overload problem of information. In the most popular microblogging website Twitter [1], hundreds of millions of people post their statuses with 140-character limit messages (which are called tweets) everyday, resulting in a very large number of online topics. Therefore, it is necessary to help users find out the interesting content in microblogging, which can be described as "which topics are interesting to user". Further, predicting users' opinions toward those topics they are interested in, which can be described as 'what opinion does user hold on a specific topic", is challenging but able to give useful feedback information for user understanding and analyzing. In this dissertation, we address above two problems as user preference predictions, and exploit the information of social context and

[1] http://twitter.com/

topical context to give solutions.

1.1 Background

Comparing with the multi-media content of comprehensive social networking services, such as Facebook, Myspace, the short text messages posted on microblogging are easier to process and analyze. As one of the most popular microblogging site, Twitter allows the users to share information with their online friends by posting 140-character limit tweets. With the growth of users and the availability of rich data resource, Twitter attracts much attention of researchers from diverse domains.

The large amount of user generated data of microblogging facilitates the study for user preference. In Twitter, the # symbol used before a word or phrase (no spaces) is called a hashtag in tweets (e.g. #iphone), to label the topics that are created organically by Twitter users. As a result, the popular hashtags are usually utilized as the topics in previous work [81, 58, 7]. In this dissertation, we also select hashtags as topics when predicting user preference. As a successful microblogging site, Twitter also offers the social network information. Mining social and topical context information to predict user preference in microblogging, presents both challenges and opportunities.

1.1.1 User Preference in Microblogging

In this dissertation, user preference is defined from two aspects: user interest and user opinion. Mining user interest in microblogging is actually to help to solve the problem "which topics are interesting to user", while inferring user opinion is to solve the problem "what opinion does user hold on a specific topic". The two tasks are described in Figure 1.1. In the remainder of this dissertation, the former one is referred to as user interest prediction, and the latter one is referred to as user opinion prediction.

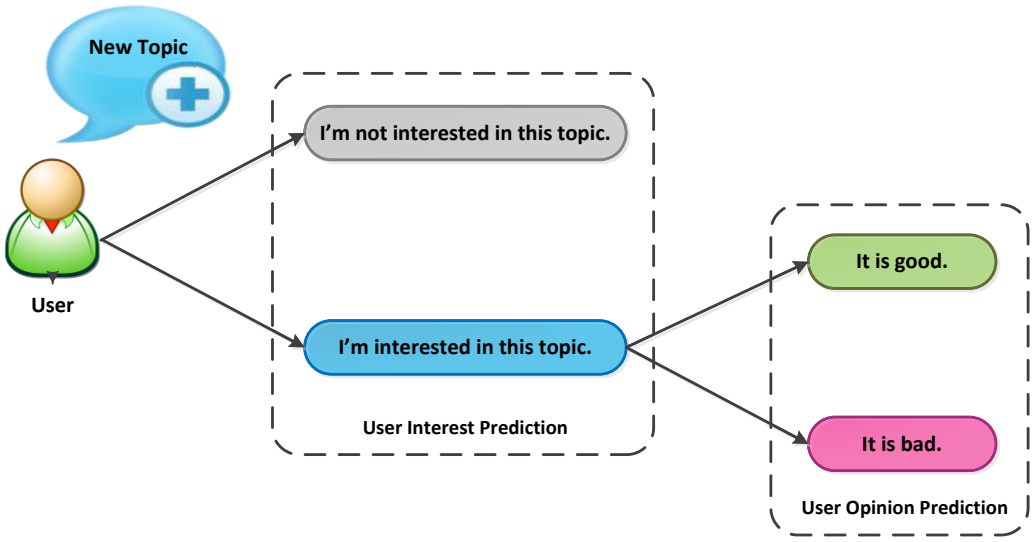


Figure 1.1: The two tasks for predicting user preference.

1.1.2 User Interest Prediction

User interest prediction is actually applied to personalized topic recommendation in microblogging. Filtering and recommending topics precisely meets users' personal information needs and save their manual effort [19], which is an important challenge. As the revenue model of microblogging is related to its huge amounts of users, improving the performance of user interest prediction contributes to make financial benefits in the real world.

In traditional recommendation tasks, the preference indication from user to item can be either explicit (such as a 1-5 scale rating) or implicit. For user interest prediction in this dissertation, users' tweeting behaviors are employed as the implicit indications. Thus if we observed that user u had posted a tweet tagged with hashtag i, but he/she had never tagged a tweet with hashtag j so far, then we would infer u was more interested in i than in j at the moment. We consider the observation as the implicit user feedback, and define it as the interest of user u to hashtag i.

Since we treat user interest prediction as topic recommendation problem, col-

laborative filtering technique that is widely used to solve this type of problems is introduced to solve it [30][96]. Previous work of recommender systems with implicit preference indications usually utilizes opinion mining techniques to analyze the content of user reviews, and assumes that the positive opinions toward item will improve its ranking in the recommendation list for user, while the negative opinions will pull down the ranking [107] [90]. However, recommending interesting topics in Twitter is different, where a topic with many positive expressions does not imply that user will have more interested in it. Actually, the user probably prefers to know those controversial topics debated by people with opposite opinions, or even those hot events bringing out a large number of negative posts. In order to infer and recommend user interest more precisely, more detailed context information is considered in our proposed solution.

1.1.3 User Opinion Prediction

For understanding user preference adequately, after predicting which topics are interesting to user, we further analyze what opinion does user hold on a specific topic, which is referred to as user opinion prediction problem in this dissertation. The high usage frequency of microblogging makes the messages posted by users are more likely to reflect their spontaneous emotions. Particularly, those subjective feelings about specific topics could be defined as users' opinions, which are considered to play an important role during their decision-making process most of the time [76]. User opinion prediction is applicable to such as viral marketing [38], opinion polling [104, 70, 117], mood monitoring [79, 11], and so on.

In the most famous microblogging Twitter, a series of research has been performed for analyzing sentiment and mining opinion from tweets [24, 17]. For a given query in Twitter, several online sentiment/opinion tracking tools also have been

developed, such as TwitterSentiment140 [2], Twitrratr [3], TweetFeel [4], etc. However, in these previous work, the researchers mainly focused on measuring the sentiment of one tweet or inferring the public opinion of mass populations, but ignored *which* is *whose* opinion.

In our work, we study what opinion user holds on a specific topic, thus to predict who has what opinion of a specific topic. This is a more difficult task but in demand in some application scenarios. For example, being aware of whether a user will like a hot product could let the company find the target customers more efficiently. During an election, detecting the individual political opinion could help the candidate know which extra portion of people may potentially be got as his/her voters. For the web sites, they could recommend the trending content what is really interesting to users if they can infer their personal tastes. Automatically mining these user-topic opinions from the user-generated and opinion-rich resource Twitter, would no doubt be an efficient and low-cost way.

Note that in all the above application scenarios, user opinion should be detected before the events occur. In the task of user opinion prediction, the most challenging problem is how to predict users' opinions towards specific topics in the case their posts have not been observed yet. In our work, the opinion homophily among social friends in microblogging, and users' opinion consistency on content-related topics, are considered and employed for predicting the unknown user opinions.

1.1.4 Exploiting Social and Topical Context

"*Birds of a feather flock together.*" The theory of *homophily* indicates that users with similar characteristics are more likely to create relationships [66]. The effect of homophily in social networks has been validated in some previous work. In Zafarani et al.'s work, they thought that sentiment/emotion may propagate through

[2]http://www.sentiment140.com/
[3]http://twitrratr.com/
[4]http://www.tweetfeel.com/

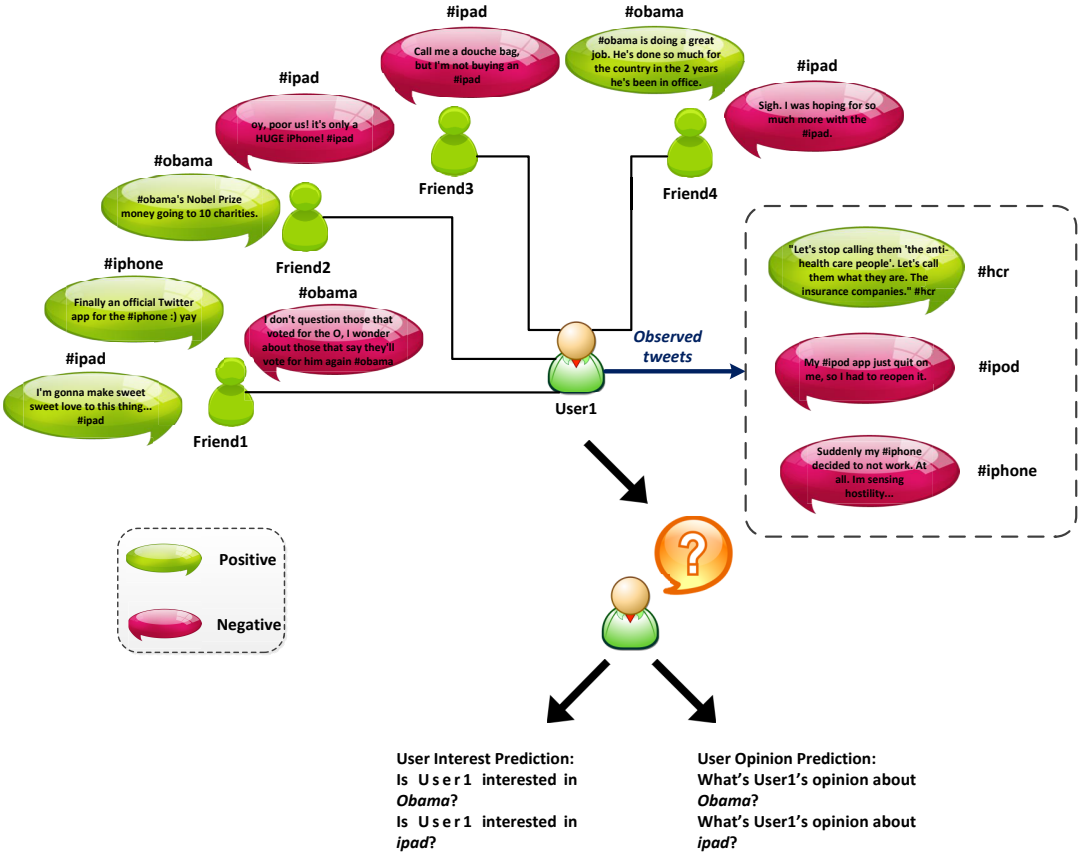


Figure 1.2: A toy example of user preference prediction.

a social network [116]. Bollen et al. showed that general happiness of users is indeed assortative across the Twitter social network [9]. A quantitative study to infer emotional states of users at a future time regarding social correlation as an important factor for prediction [100]. Tang et al. demonstrated the existence of homophily in trust relationship network [98]. According to the homophily theory, we define the social network information as social context, and exploit it for the user preference prediction tasks.

Like the relationships among users, there also exist correlations among topics. In previous related work, topic correlation was exploited to help identify whether two citations with the same author name refer to the same individual in [113]. Lai

and Liu computed the topic similarities of features, and took them as a factor when evaluating the orientation of texts [37]. Based on the assumption that the topics more similar will be given more similar interests/opinions, we seek to infer user preference on a specific topic according to the existing observations about those closely related topics. The extracted topic correlation information is defined as topical context in this dissertation.

Figure 1.2 gives a toy example illustration about user preference prediction tasks taking the social and topical context into consideration. In this example, four social friends of User1 posted tweets to talk about how they feel about *Obama*, *ipad* and *iphone* in Twitter. User1 is observed that he/she gave his/her opinion on health care reform (*hcr*), and murmured at his/her *ipod* and *iphone* in the past. Exploiting the social context and topical context, we first attempt to predict whether he/she will be interested in the topics of *Obama* and *ipad*. If so, we then further predict whether he/she will support *Obama*, and whether he/she will like the *ipad*.

Aiming to incorporate the information of social and topical context, we model the user preference prediction tasks with collaborative filtering techniques, and finally propose the Social context and Topical context incorporated Matrix Factorization (ScTcMF) framework to achieve the goal. This framework is quite general, which can be easily applied to both user interest prediction and user opinion prediction. A real-world dataset is collected from Twitter for evaluation. In the experiments on real-world Twitter dataset for the two user preference prediction tasks, the proposed ScTcMF framework is compared with the state-of-the-art collaborative filtering methods. The experimental results demonstrate that the ScTcMF framework with social and topical context leads to improvements in both two user preference prediction tasks, even when the observed training data is sparse.

1.2 Contributions

In this paper, we investigate how the social and topical context information can help enhance the user preference prediction tasks. The hypotheses based on real-world data observation and analysis are formulated in user interest prediction and user opinion prediction, respectively. Then they are mathematically modeled, and employed by a proposed social and topical context incorporated framework. To the best of our knowledge, this is the first work that social and topical context information is combined for predicting user preference in microblogging.

The main contributions of this dissertation are:

- We propose a general framework for incorporating social context and topical context as regularization constraints to help improve the performance of two user preference prediction tasks in microblogging: user interest prediction and user opinion prediction.
- For predicting user interest, we first exploit the characteristics of topical opinion distribution to describe topical context information, and further capture the weights between social friends under different opinion distribution topic patterns as social context information.
- For predicting user opinion, we utilize social friend relationships between users as social context information, and content-based correlations among topics as topical context information.
- The proposed framework is empirically evaluated on a real-world Twitter dataset, and the experimental results exhibit its good performance.

1.3 Organization

The remainder of this dissertation is organized as follows. We first give a brief literature review in Chapter 2. Chapter 3 gives the description of Twitter data collection

and preprocessing in detail. Chapter 4 introduces the low-rank matrix factorization method as the basic model, and describes how to mathematically incorporate social context and topical context by developing regularization constraints. Finally ScTcMF framework is presented in this chapter. In Chapter 5 and Chapter 6, we investigate the effect of social and topical context on the tasks of user interest prediction and user opinion prediction, respectively. In Chapter 5, we experimentally apply the proposed ScTcMF framework to user interest prediction, and evaluate the performance of ScTcMF and compare against the state-of-the-art methods. In Chapter 6, empirical results about ScTcMF framework for user opinion prediction are reported. We conclude the dissertation and point out future research directions in Chapter 7.

Chapter 2

Literature Review

In this dissertation, we propose a framework incorporating social and topical context for predicting user preference in microblogging. User preference is defined from two aspects: user interest and user opinion. Correspondingly, the two tasks of our study are user interest prediction, and user opinion prediction. There have been a lot of previous work related to these two tasks and inspire our work. In the following sections, we will give a literature review on them respectively.

2.1 Recommender Systems

User interest prediction actually equates to personalized topic recommendation in microblogging. In this section, we first review some main techniques of general recommender systems, containing both non-personalized approaches and personalized approaches. The second subsection concentrates on the review of personalized recommender systems. At last, existing work about personalized recommendation in microblogging is presented.

2.1.1 Techniques of General Recommender Systems

Recommender systems refer to technologies seeking to predict the rating or preference that user would give to an item [85]. Recommender systems have been applied in a variety of applications, which can be generally classified into three categories according to the method to produce a recommendation list: collaborative filtering, content-based filtering, and hybrid recommender systems [2].

Among these three categories, collaborative filtering is the most widely used technique to build recommender system [30, 96]. The typical collaborative filtering algorithms empirically learn a dataset of preferences to recommend appropriate items to users. Given a list of users and a list of items, the past behaviors of users can be analyzed to infer their potential preferences. Usually, the preference from user to item is either explicit indication in those traditional 5-star rating systems, or implicit indication such as click-through [69], location check-in [22], trust relation [98] or other user behavior [14, 73, 19]. In our task of user interest prediction in microblogging, we employ users' tweeting behaviors as the implicit indications, due to the absence of high quality ratings in microblogging.

Further, collaborative filtering methods can be further classified into memory-based collaborative filtering and model-based collaborative filtering, and hybrid collaborative filtering.

Most of early collaborative filtering systems use memory-based methods, which infer preferences according to the calculated similarities between the neighbors. The memory-based collaborative filtering methods are simple but effective, and have been adopted in many applications. According to whose similarity it relies on to perform the recommendation, memory-based methods contain user-based [29] and item-based [87].

The user-based method is the most common form of memory-based collaborative filtering [48, 12, 41]. The idea of user-based method is to capture a user u's preference on unobserved items based on the preferences from K users most simi-

lar to him/her. Herlocker et al. analyzed design choices of user-based collaborative filtering algorithms in their experiments. They divided the neighborhood-based prediction approach into three components identified as similarity computation, neighbor selection, and rating combination. Analogously, the item-based methods firstly finds K most similar items and then calculates a weighted average of their observations. Amazon.com used item-item collaborative filtering method to produce a list of product recommendations for each customer [59]. Different techniques for computing item-item similarity are investigated in [87]. In the paper of Ma et al., they presented an algorithm to predict the missing data with a combination of user and item information. Their algorithm is also able to determine whether to predict the missing data or not [61].

Although memory-based collaborative filtering methods are efficient and easy to adopt, there are several disadvantages with them. In particular, the whole user-item matrix employed by memory-based collaborative filtering methods is usually very sparse in many real-world large datasets. Under sparse data, the similarity measured from ratings/preferences may not be reliable due to the insufficient information observed [78], which decreases the recommendation performance and prevents the scalability of memory-based methods.

Model based collaborative filtering methods leverage data mining and machine learning technologies to learn a model from training data, and applies the model on test data to predict user preferences on different items. Various collaborative filtering models are investigated, including clustering models [106, 112], latent factor models [33, 32], etc.

A clustering collaborative filtering model based on hierarchical clustering is presented by [47]. Hofmann proposed a latent factor model based on a generalization of probabilistic latent semantic analysis to continuous valued response variables [31]. Among the latent factor models, the matrix factorization model has been widely used in recent years.

Matrix factorization becomes a successful collaborative filtering technique due to its scalability, flexibility, and the predictive accuracy [53, 49, 108]. The basic idea of the matrix factorization is to assume that there are certain latent factors related to both the preferences of users and the properties of items. An algorithm of weighted low-rank matrix factorization approximations is proposed by Srebro et al., and applied to collaborative filtering task [92]. The probabilistic matrix factorization model which scales linearly with the number of observations is described in [86]. Koren et al. demonstrated that the flexibility of matrix factorization framework allows the incorporation of additional knowledge [49]. Gu et al. proposed a matrix factorization model incorporating user and item graphs [26], which is inspiring for our work.

In order to improve the recommendation performance, some researchers also proposed hybrid methods by combining both collaborative and content information [82, 54]. The hybrid collaborative filtering methods overcome the limitations in native collaborative filtering methods such as sparsity and loss of information, but they also increase algorithm complexity and are expensive to implement [23]. The hybrid methods are adopted in most of the commercial recommender systems. Google news recommender system is one famous example [16].

2.1.2 Recommender Systems Using Context Information

Existing work employed various context information to offer more precise recommendation. In [3], Adomavicius and Tuzhilin investigated the effect of relevant context information in recommender systems, and showed that it is important to take context information into account when providing recommendations. By adding the contextual information, such as when, where and with whom a movie is seen, memory-based collaborative filtering method could outperform the pure traditional method without any additional method in movie recommendation application [1]. A music recommender system is proposed in [95], which tackles the music recommen-

dation problem by mining musical content and context information. In the work of Lu et al., they exploit context information about authors' identities and social networks for improving review quality prediction [60].

With the availability of the information about user relationships in the social networks, the concept of social recommendation was proposed in [62], the authors believed that exploiting a users social network graph would make more accurate and personalized recommendations. Specially, the *trust* relationships between online users are considered to play an important role in product recommendation. Trust-aware recommendation systems were thus investigated by recent research [65, 63, 99]. Ma et al. provided a general method that can be utilized to both social recommender systems and trust-aware recommender systems [63]. Tang et al. studied the multi-faceted trust relationships between users, and incorporated these relationships into rating prediction [99].

With the growing availability of opinion-rich resources online, some researchers regard the opinions expressed in user reviews as important external content information to help recommendation. Zhu et al. presented an aspect-based opinion polling algorithm based on the data of Chinese restaurant reviews [117]. They also proposed an aspect-based segmentation algorithm for restaurant rating inference [118]. Wang and Chen built product reviewers' preference similarity network considering their opinion values on features [107]. Stavrianou and Brun used opinions extracted from user reviews as fine-grained information to improve an expert recommender system [93]. Sohail et al. presented a book recommender system using opinion mining technique to propose top ranked books [90]. These papers are mainly based on the assumption that the positive opinions toward item will improve its ranking in the recommendation list, while the negative opinions will pull down its ranking. In the task of user interest prediction in this dissertation, we exploit topical opinion distribution characteristics rather than opinion expression weights to help predict interesting topics for users.

2.1.3 Personalized Recommendation in Microblogging

The problem of recommending valuable information for users in microblogging has attracted increasing attention. In this subsection, we mainly introduce some related work about personalized recommendation in Twitter, as it is one of the most popular microblogging services.

Content recommendation on Twitter was empirically studied in [13]. For filtering information stream of Twitter, Kapanipathi et al. proposed an architecture to filter and deliver interesting tweets to users [43]. Hong et al. investigated the problem of predicting the popularity of messages measured by the number of future retweets, which is helpful for the task of personalized message recommendation [34]. To prevent users from the information overload problem, Chen et al. gave a solution of personalized tweet recommendation based on collaborative ranking [14]. Pan et al. solved the problem by proposing a framework integrating both the advantages of collaborative filtering and the characteristics of diffusion processes later [73].

Note that the goal of the above research is to recommend tweets rather than topics to users. Aiming to discover the topics of interest for Twitter users, Michelson and Macskassy proposed an entity-based profiling approach, which leverages a knowledge base to disambiguate and categorize the entities in tweets [68]. The topic discovery and recommendation in Twitter was also addressed in the work of Diaz-Aviles et al [19].

Besides, some other recommendation tasks were conducted and estimated on Twitter. Hannon et al. built a followee recommender system for Twitter users using content and collaborative filtering approaches [27]. Later, they evaluated a variety of different recommendation strategies for finding useful users on Twitter [28]. In [45], Kim et al. proposed a recommendation system named TWITOBI for Twitter. With a probabilistic model utilizing not only tweet messages but also the relationships between users, TWITOBI can recommend top-K users to follow and top-K tweets to read for a user.

2.2 Research about Opinion Prediction

The task of predicting users' opinions toward specific topics they had not directly given is challenging, and different from most of existing work. In this section, we mainly review research related to the user opinion prediction, including sentiment analysis and opinion mining techniques, microblogging data based sentiment analysis, and applications using opinions detected from microblogging.

2.2.1 Sentiment Analysis and Opinion Mining

Sentiment analysis and opinion mining refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials [1]. Generally, it aims to detect the sentiment/opinion polarity (positive, negative, or neutral) of a given text at the document, sentence, or feature/aspect level.

Given a piece of text, the subtasks of sentiment analysis and opinion mining mainly include [75, 46]:

- which part is sentiment/opinion expressing;
- who wrote the sentiment/opinion;
- what is being commented on;
- what is the sentiment/opinion of the writer.

Many previous works have been proposed to investigate sentiment analysis and opinion mining. Most of early work in this domain employed lexicon-based approaches to analyze sentiment/opinion in text [44, 21, 42]. MPQA subjectivity cues lexicon is well-known resource developed by researchers in University of Pittsburgh [111, 110]. There are some other publicly available resources that can be used to

[1] http://en.wikipedia.org/wiki/Sentiment_analysis

extract the semantic and affective information associated with natural language concepts for building systems of sentiment analysis and opinion mining [94, 5].

Pang et al. applied machine learning techniques to conduct document-level sentiment classification, and compare the classification performance of different machine learning models [77]. Using the Pointwise Mutual Information and Information Retrieval (PMI-IR) algorithm to estimate the semantic orientation of the extracted phrases, Turney presented an unsupervised learning algorithm for classifying the opinions expressed in product reviews [105]. Pang and Lee used an efficient and intuitive graph-based formulation relying on finding minimum cuts to extract the subjective portions of the document, and then applied text-categorization techniques to just the subjective portions for sentiment analysis [74].

Instead of classifying the sentiment of an entire document, Yi et al. presented a sentiment analyzer that detects sentiments about a given topic using natural language processing techniques [114]. Mei et al. defined the problem of topic-sentiment analysis, and propose a novel probabilistic model to capture the mixture of topics and sentiments simultaneously [67]. However, they did not model sentiment directly, and their model required post-processing to identify the polarity of a document. In [56], Lin and He proposed unsupervised joint sentiment/topic mode based on Latent Dirichlet Allocation (LDA) to detect sentiment and topic simultaneously.

2.2.2 Sentiment Analysis in Microblogging

The rise of microblogging in Web 2.0 age has fueled interest in sentiment analysis in the past years. In this subsection, we introduce existing work of sentiment analysis in microblogging.

In some early work, the researchers applied the state-of-the-art sentiment analysis and opinion mining methods designed for traditional long text data to the text in microblogging. Go et al. used distant supervision to automatically classify the tweets as positive or negative sentiment [24]. The main contribution of their paper

Table 2.1: List of emoticons verified by Twitter API.

Emoticons mapped to :)	Emoticons mapped to :(
:)	:(
:-)	:-(
:)	: (
:D	
=)	

is the idea of using tweets with emoticons for distant supervised learning. They employed the emoticons verified by Twitter API expressing positive emotion and negative emotion as labels to train classification models. The full list of emoticons that they used can be found in Table 2.1. The performances of different machine learning models for classifying tweet sentiment were then compared in this paper.

Davidov et al. proposed a supervised framework by further utilizing 50 Twitter tags and 15 smileys as sentiment labels, and evaluated the contribution of different feature types for sentiment classification [17]. Pak and Paroubek showed how to automatically collect a Twitter corpus for sentiment analysis and opinion mining purposes, and they built a multinomial Naïve Bayes classifier to determine positive, negative and neutral sentiments for tweets [71]. Barbosa and Feng proposed a 2-step sentiment analysis classification method, which first classified tweets as subjective and objective, and further distinguished the subjective tweets as positive or negative [6]. They validated that the proposed method is robust to the noisy and biased Twitter data in their paper. In [50], the authors evaluated the usefulness of existing lexical resources, as well as features that capture information about the informal and creative language used in microblogging, and finally trained supervised models to mine Twitter sentiment about given topics.

In recent papers, researchers started to take the characteristics of Twitter into account, and proposed novel approaches for sentiment analysis in Twitter. In the work of Speriosu et al. [91], they proposed a label propagation approach to make polarity classification for tweets, and exploited the Twitter follower graph to assist

in the classification. In Jiang et al.'s paper [40], in order to improve tweet sentiment classification, they incorporated target-dependent features, and took relationships between tweets into consideration. However, they just tested on a small size dataset, and their sentiment analysis is just at tweet-level but not at user-level. Hu et al. presented a supervised learning method to investigate whether social relations can help sentiment analysis [36]. Tan et al. also proposed a user-level sentiment analysis model exploiting social network information in Twitter [97]. They collected data produced by groups of extremely opinionated users to evaluate their model. Their final data contained five selected topics, and the correlations between topics weren't considered in their study.

2.2.3 Detecting Opinion from Microblogging for Applications

The opinion and emotion detected from Twitter have also been exploited for applications in various domains. Considering the online mention of a brand plays an important role in customer buying decisions, Jansen et al. reported a study investigating Twitter as a form of electronic word-of-mouth for brand management [38]. They analyzed more than 150,000 tweets containing branding comments, sentiments, and opinions, and compared automated methods with manual coding for classifying sentiment in those tweets.

Based on the assumption that opinions in social media correlate to what happened in the real world, O'Connor et al. linked sentiment of text on Twitter to public opinion from traditional polling data on consumer confidence and political opinion [70], and they found their sentiment detector based on Twitter data replicated those poll data from traditional survey methodology to an extent, and could be considered as a substitute for traditional polling. Tumasjan et al. conducted a content analysis of over 100,000 messages containing a reference to either a political party or a politician, and their results also validated that the activity on Twitter

can be used to predict the popularity of parties or coalitions in the real world [104]. Skoric et al. sought to forecast the election results of the 2011 Singapore General Election, using Twitter data obtained during the official campaign period [89].

The investigation results of Bollen et al. showed that using Twitter data could model public opinion and emotion, and had a predictive power for socioeconomic phenomena [11]. Another research of them employed public opinion expressed in Twitter posts to predict the trend of the Dow Jones Industrial Average (DJIA) [10]. Their results indicated that the accuracy of predicting the daily up and down changes in the closing values of the DJIA is more than 80%.

For studying the spread of bad news through social media, Park et al. designed a case study on the Dominos Pizza crisis in 2009, by analyzing the sentiments of related tweets [79]. Golder and Macy's study utilized data from Twitter to identify individual-level diurnal and seasonal mood rhythms in cultures across the globe [25]. Asur and Huberman utilized sentiments extracted from Twitter to help forecast box-office revenues of movies [4].

Chapter 3

Data Collection and Preprocessing

Before describing the proposed framework for predicting user preference in microblogging, in this chapter we describe the real-world dataset collection from Twitter, and introduce some preprocessing implemented on the dataset.

3.1 Data Collection

In this subsection, we present why we choose Twitter as our experimental data, and describe how we collect data from Twitter using API.

3.1.1 Why Twitter

Microblogging has become a staple for users in the age of Web 2.0, which provides users the ability of exchanging information with each other, in a more simple and convenient way. As well known, Twitter is the most famous and popular microblogging service site. After created in March 2006, Twitter service that enables users to send and read short 140-character messages, rapidly gained worldwide popularity. Within a few months of its launch, Twitter had about 94,000 users as of April, 2007 [1]. In 2012, it has already earned more than 100 million users who posted 340

[1] http://www.usatoday.com/tech/webguide/2007-05-28-social-sites_N.htm

Figure 3.1: An example of Twitter homepage.

million tweets per day [2]. Figure 3.1 [3] shows a snapshot of Twitter homepage. The wide popularity is no doubt an important reason why we choose Twitter as the experimental data in this dissertation.

Twitter provides a simpler and faster mode of communication, compared to traditional blogging service. Sending and reading those 140-character limit messages on Twitter called *tweets*, saves time for users, and encourage them to generate fresh new content and exchange information more frequently. In traditional blogging, a user may post a blog every few days. However, several tweets may be posted every day in Twitter [39, 52].

Compared to the content in those comprehensive social networking services, the content of tweets are easier to obtain and analyze. Because all we need to process

[2]https://blog.twitter.com/2012/twitter-turns-six
[3]https://about.twitter.com/press/accounts

are the short 140-character limit messages, but not multiple types of data.

3.1.2 Crawling Data with Twitter API

For collecting data from Twitter, we utilized the set of APIs offered by Twitter [4] for developers. With the basic breadth-first search strategy, we started with a set of active users, and further found their friends and followers in the social networks of Twitter.

For each user, we collected information about profiles, and the tweets they posted from their Twitter account were created to *January 2011*. We crawled data from Twitter until that a collection of $10,000$ users' profiles and the tweets are obtained. The detailed profiles of the collected Twitter Dataset are shown in Table 3.1.

Table 3.1: The profiles of Twitter Dataset.

Item	Description
user_id	It is a string of numbers to identify unique user in Twitter.
realname	It is a personal identifier displayed in user's profile page.
username	It is used for logging in and is unique.
location	It shows the location setting by user.
description	It is the self-description written by user.
url	It is a unique vine profile address accessible from the web.
followers_count	It is the number of user's followers recently.
friends_count	It is the number of user's friends recently.
created_time	It shows the time when user created the Twitter account.
favorites_count	It is the number of user's favorites recently.
time_zone	It shows the time zone of user.
geo_enabled	It shows the location where user is tweeting, valid only if user enabled location services.
verified	It is used to detect if a user is verified on Twitter.
statuses_count	It is the number of the statuses posted by user recently.
lang	It shows the language setting by user.
friends	It lists all the friends of user.
followers	It lists all the followers of user.

[4] http://dev.twitter.com

3.2 Data Preprocessing

In this subsection, we introduce the preprocessing implemented on Twitter data for the user preference prediction experiments, mainly including: language selection, user filtering, and user-topic opinion labeling.

3.2.1 Language Selection

Since the users of Twitter mainly live in the English-speaking regions, we selected those users who set the language option in their profiles as "en" for the user preference prediction tasks. After this preprocessing, we found there were a portion of users with "en" language option in their profiles actually posting non-English tweets in the dataset. Therefore, we also employed Microsoft Translator API [5] to filter the non-English content.

3.2.2 User Filtering

After the processing of language selection, there are still $8,705$ users' profiles and their tweets in our collected Twitter dataset. These users are different in their activities. Some of them updated few tweets since their Twitter account were created. This portion of users is hardly to learn useful information in both of our two user preference prediction tasks. Therefore, we filter the inactive users when we carry out experiments in the prediction tasks. The datasets selected for user interest prediction and user opinion prediction are different. We will describe them in detail in Chapter 5 and Chapter 6, respectively.

3.2.3 User-topic Opinion Labeling

In this dissertation, the set of labeled user-topic opinions is needed in the process of modeling for both user interest prediction and user opinion prediction. However,

[5] http://msdn.microsoft.com/en-us/library/ff512423.aspx

due to the huge amount of tweets in our dataset, labeling the opinions manually is very costly and impractical. Hence we need an automatic approach to label the user-topic opinions.

In some previous papers about sentiment analysis on Twitter, the authors usually use specific marks in tweets, such as opinion keywords/phrases or emoticons to assign sentiment labels [24, 71]. In this work, we adopt a tool named *SentiStrength* [6] to label the user-topic opinions in Twitter dataset [103, 102].

SentiStrength is built for estimating the strength of positive and negative sentiment in short social web text in particular, which reports human-level accuracy and has been applied in some related research [101, 80, 51]. It develops a sentiment lexicon containing opinion keywords as well as emoticons and slang in the web, and combines the lexicon based approach with some linguistic rules, such as negation detection and spelling correction. Table 3.2 lists the main lexicons and linguistic rules employed by SentiStrength [102]. Figure 3.2 shows a snapshot of quick test given by SentiStrength.

Table 3.2: The list of main lexicons and linguistic rules.

Name	Description
Sentiment word list	It is a word list with human polarity and strength judgements.
Spelling correction	It deletes repeated letters in a word when the letters are more frequently repeated than normal.
Booster word list	It is used to strengthen or weaken the emotion of following sentiment words.
Idiom list	It is used to identify the sentiment of common phrases.
Negating word list	It is used to invert following emotion words.
Repeated letters	At least two repeated letters added to words give a strength boost sentiment words by 1.
Emoticon list	It lists the emoticons with polarities to identify additional sentiment.
Repeated punctuation	One or more exclamation marks boost the strength of the preceding sentiment word by 1.

[6] http://sentistrength.wlv.ac.uk/

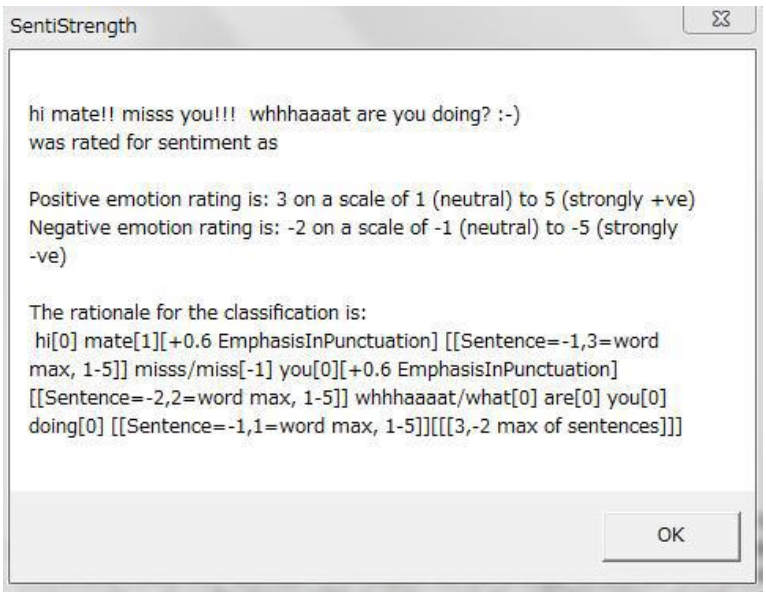


Figure 3.2: A quick test example of SentiStrength.

For each text, SentiStrength reports two integers: a positive strength score ps ranges from 1 (not positive) to 5 (extremely positive), and a negative strength score ns ranges from -1 (not negative) to -5 (extremely negative). Let $Tw(u, i)$ denote the text of all tweets posted by user u about topic i. We expect SentiStrength to label user-topic opinion $O(u, i)$ for $Tw(u, i)$. A simple and intuitive labeling method is applied: $O(u, i)$ is labeled as 1 if $ps + ns$ is positive, and -1 if $ps + ns$ is negative. User u is regarded as neutral on hashtag i if $ps + ns$ equals to 0.

Chapter 4

Social and Topical context incorporated Framework

In this chapter, we propose the Social context and Topical context incorporated Matrix Factorization (ScTcMF) framework for predicting user preference in microblogging. We first present low-rank matrix factorization as the basic prediction model, and then interpret how to mathematically incorporate social context and topical context, respectively.

4.1 The Basic Low-rank Matrix Factorization Model

Due to the predictive accuracy, scalability and flexibility for incorporating additional information, matrix factorization methods are widely employed in the state-of-the-art collaborative filtering tasks [84, 49, 26, 63]. In Su et al.'s survey of collaborative filtering techniques, they explain that traditional CF algorithms will suffer serious scalability problems in recommender systems as numbers of existing users and items grow, but matrix factorization is a technique of dimensionality reduction which can deal with the scalability problem and quickly produce good quality recommendations [96]. Koren et al. regard that the flexibility of matrix factorization allows incorporation of additional knowledge as one of its important strength [49].

Let $\mathbf{u} = \{u_1, u_2, ..., u_m\}$ be the set of users, and $\mathbf{t} = \{t_1, t_2, ..., t_n\}$ be the set of topics, where m and n are the numbers of users and topics, respectively. A user-topic matrix $M \in \mathbb{R}^{m \times n}$ consists of element $M(u, i)$, which represents the preference of user u on topic i. In our case, since the observed data in the real-world Twitter dataset is only a small percent, the user-topic matrix M is very sparse. Therefore, on the premise that only a small number of factors influence the preferences [84], we give a more compact but accurate representation for users and topics in a low-rank space, and attempt to approximate the matrix M by a multiplication of low-rank factors, as the following:

$$M \approx UH^T \tag{4.1}$$

where $U \in \mathbb{R}^{m \times d}$ and $H \in \mathbb{R}^{n \times d}$ with $d \ll min(m, n)$, The row vector $U(i, :), 1 \leq i \leq m$ in U, and $H(j, :), 1 \leq j \leq n$ in H are the latent representations of user i and topic j in low-rank space respectively. The matrix factorization method traditionally approximates the matrix M by minimizing the following objective,

$$\min_{U,H} \|M - UH^T\|_F^2 \tag{4.2}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix, and $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |A(i, j)|^2}$. Because M contains a mass of unknown elements, we introduce an indicator matrix $Y \in \mathbb{R}^{m \times n}$ to only model the observed data, in which $Y(u, i) = 1$ if user u express his/her preference on topic i and $Y(u, i) = 0$ otherwise. Additional regularization terms on U and H are added to avoid overfitting, as suggested by some recent works [49]. Hence we have

$$\min_{U,H} \|Y \odot (M - UH^T)\|_F^2 + \lambda_1 \|U\|_F^2 + \lambda_2 \|H\|_F^2 \tag{4.3}$$

where the symbol $\odot$ in the equation is Hadamard product, by which $(A \odot B)(i, j) = A(i, j) \times B(i, j)$. To avoid over-fitting, two smoothness regularizations are also added

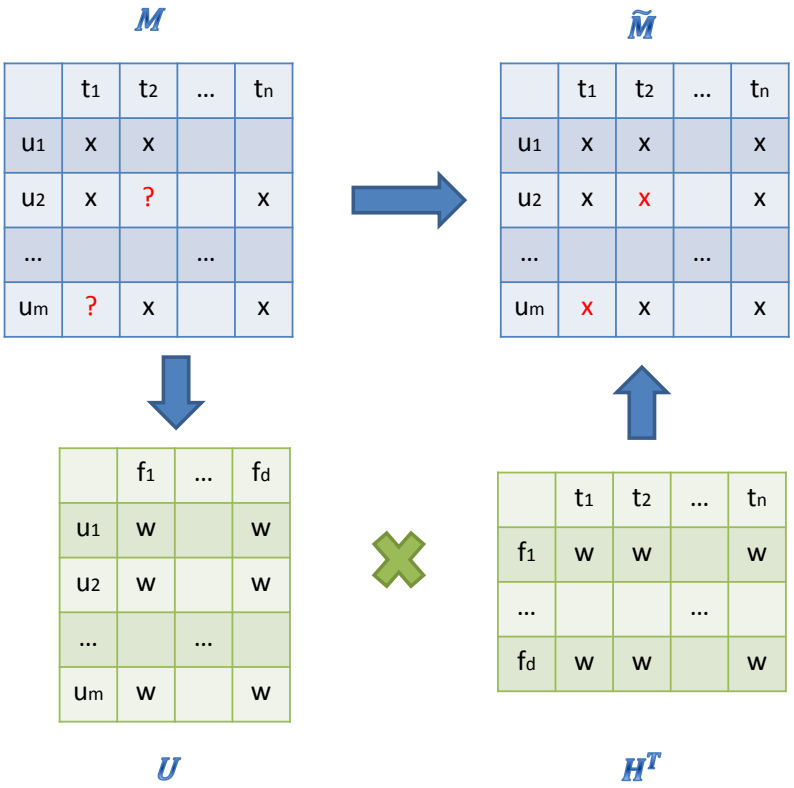


Figure 4.1: The process representation of basic low-rank matrix factorization.

on U and H. $\lambda_1, \lambda_2 > 0$ are the smoothness parameters to control the capability of U and H, respectively.

What is presented in Eq. 4.3 is a basic low-rank matrix factorization model. The process representation of basic low-rank matrix factorization is illustrated in Figure 4.1. There have been many existing approaches can find a optimal solution for it [15, 55, 115]. In the following sections, we will discuss how to incorporate social and contextual context into this basic matrix factorization model.

4.2 Social Context Regularization

We present the definition of social context, and the social context hypothesis in microblogging in detail. Using the proposed hypothesis, a regularization constraint

term is formulated for user preference prediction. This regularization constraint describes how to utilize the relationships between users in social networking, and take them into account when construct the objective function.

4.2.1 Social Context Hypothesis

Like most of social networking services, microblogging allows users to create explicit relationships with others. The users with social relationships usually exchange their information online. Those observed microblogging users and the social relationships created by them provide a social context for user preference prediction. In this subsection, we formally give the definition of *social context* as follows.

Definition 1 (Social Context) *Social context is defined as a graph* $G_S = \{\mathbf{u}, \mathcal{S}\}$ *with adjacency matrix* S*. The non-diagonal element* $S(i, j)$ *in* S *is a weight value within the range* $[0, 1]$ *if user* i *created a social relationship with user* j*. The rest of non-diagonal elements and the diagonal elements in* S *are set to* 0*.*

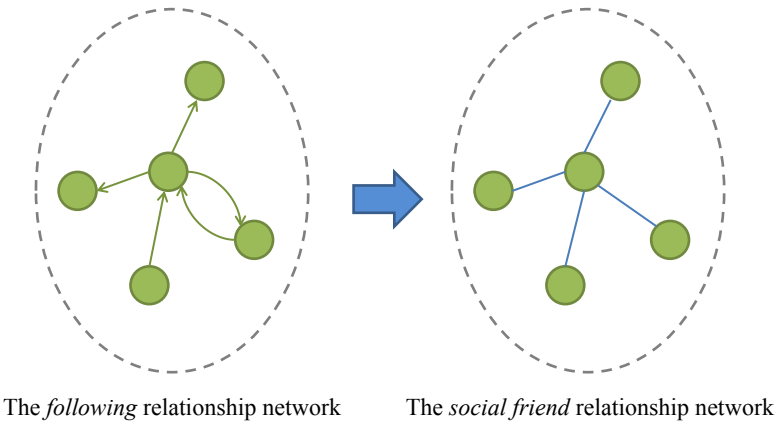


Figure 4.2: *Social friend* relationship network in social context hypothesis.

As the chosen microblogging service for empirically evaluation, Twitter provides a *following* mechanism enabling a user to follow any other users. A user who follows other users is called as their *follower*. A user who is followed by other users is called as the followers' *friend*, no matter whether they follow back or not. Following

friends on Twitter means that the follower is subscribing to their tweets, and the updates of friends will appear in the follower's *Home* tab. Except some special cases (e.g. celebrity following), creating a following relationship usually implies that the follower and the friend may have more similar preferences towards the same topic than those non-friends without any explicit relationship, with higher probability. Therefore, for effective formulation, we define both a user's friends and followers as his/her *social friends*. The original directed following relationship network in Twitter is thus converted into an undirected *social friend* relationship network for formulating the social context in this dissertation, as shown in Fig. 4.2.

Note that after the conversation of relationship network, the social context in Twitter can be constructed as an undirected weighted graph, with a symmetric adjacency matrix S. Next, according to the homophily theory we have in introduced in subsection 1.1.4, we describe the social context hypothesis as the following:

Hypothesis 1 *With high probability, the social friends hold more similar preferences on the topics than the non-friends.*

This hypothesis is a general hypothesis about social context. In Chapter 5 and Chapter 6, we will propose specific hypotheses based the general one, for user interest prediction task and user opinion prediction task respectively, and validate them experimentally on the real-world dataset.

4.2.2 Exploiting Social Context for Regularization

Based on the above hypothesis, we consider the relationships between social friends to improve the basic matrix factorization model. The process representation of matrix factorization exploiting social context is illustrated in Figure 4.3.

Given a pair of social friends i and j, we are able to define the weight $S(i, j)$ between them depending on different prediction tasks. The definitions will be presented in Chapter 5 and Chapter 6. With the defined weights between social friends,

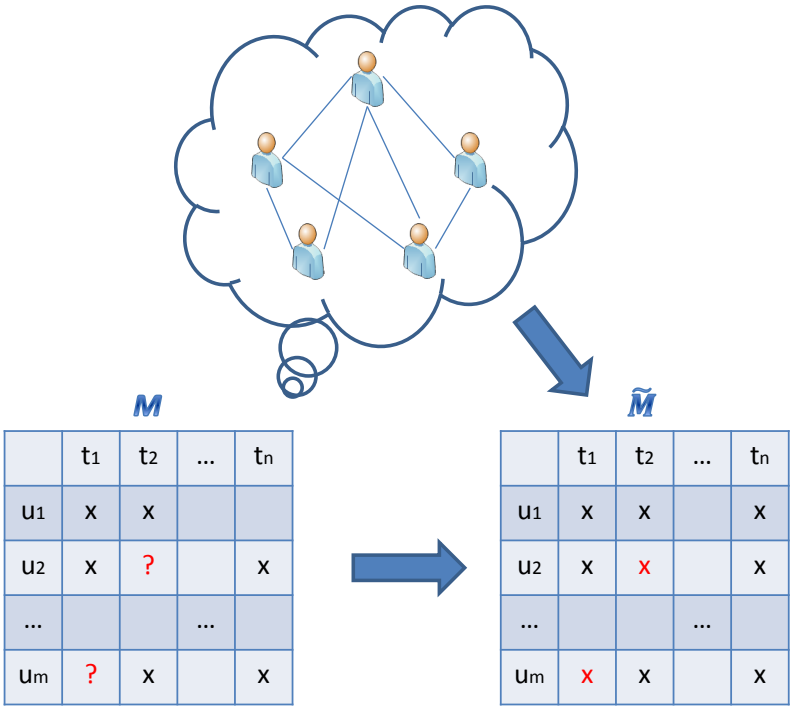


Figure 4.3: The process representation exploiting social context.

we propose a social context regularization to minimize the following terms:

$$\min \sum_{i=1}^{m} \sum_{j \in \mathcal{F}(i)} S(i,j) \|U(i,:) - U(j,:)\|_F^2 \tag{4.4}$$

In this equation $\mathcal{F}(i)$ denotes the set of social friends of user i. A small value of weight $S(i,j)$ allows larger divergence of opinion between i and j, while a large value of weight $S(i,j)$ indicates the divergence between i and j should be smaller. This regularization models a particular user and his/her friends individually, which makes the latent representation more accurate. It also has an advantage that indirectly models the propagation in the network graph of users [63].

After some derivations, we can get the matrix form of Eq. 4.4,

$$
\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{m}\sum_{j\in\mathcal{F}(i)} S(i,j)\|U(i,:)-U(j,:)\|_F^2 \\
&=\frac{1}{2}\sum_{i=1}^{m}\sum_{j\in\mathcal{F}(i)}\sum_{k=1}^{d} S(i,j)(U(i,k)-U(j,k))^2 \\
&=\sum_{i=1}^{m}\sum_{j\in\mathcal{F}(i)}\sum_{k=1}^{d} S(i,j)U^2(i,k)-\sum_{i=1}^{m}\sum_{j\in\mathcal{F}(i)}\sum_{k=1}^{d} S(i,j)U(i,k)U(j,k) \\
&=\sum_{k=1}^{d} U^T(:,k)(D_S-S)U(:,k) \\
&=Tr(U^T L_S U). \qquad (4.5)
\end{aligned}
$$

In the above equations, $Tr(\cdot)$ denotes the matrix trace, D_S is a diagonal matrix with the ith diagonal element $D_S(i,i) = \sum_{j=1}^{m} S(i,j)$, and $L_S = D_S - S$ is the Laplacian matrix.

The matrix factorization model incorporating social context regularization can be formulated as:

$$\min_{U,H} \|Y \odot (M - UH^T)\|_F^2 + \lambda_1\|U\|_F^2 + \lambda_2\|H\|_F^2 + \alpha Tr(U^T L_S U) \qquad (4.6)$$

where $\alpha \geq 0$ is a regularization parameter balancing the reconstruction error between the social context regularization term and the front terms.

4.3 Topical Context Regularization

In this section, we present the definition of topical context, and discuss how to model a topical context regularization constraint for user preference prediction. In the tasks of user interest prediction and user preference prediction, we respectively exploit different topical information for formulating topical context hypotheses, and enforce them by adding corresponding regularization constraints.

4.3.1 Topical Context Hypothesis

As mentioned in the part of introduction, the correlations among different topics are also considered to be helpful for predicting the unknown user preferences towards topics in microblogging. To capture the topic correlation information, we have the following definition of *topical context*.

Definition 2 (Topical Context) *Topical context is defined as a graph* $G_T = \{\mathbf{t}, \mathcal{T}\}$ *with adjacency matrix* T*. The non-diagonal element* $T(i, j)$ *in* T *is a value within the range* $[0, 1]$ *to weight the correlation between two different topics* i *and* j*. The diagonal elements in* T *are set to* 0*.*

Different from the social friend relationships between users in social context, there are no explicit links between topics, so we exploit different topical information for formulating topical context hypotheses for different tasks. In the task of user interest prediction, we employ the opinion distribution similarity between topics to describe their correlation. In the task of user opinion prediction, we employ the content-based correlations between topics for predicting the user-topic opinions.

In general, the hypothesis modeled for topical context is as the following:

Hypothesis 2 *With high probability, two topics more similar will be given more similar preferences by the users.*

In Chapter 5 and Chapter 6, we will propose specific hypotheses based the general one respectively, and validate them experimentally on the real-world dataset.

4.3.2 Exploiting Topical Context for Regularization

In this subsection, we consider incorporating the information of topical context based on the hypothesis in Subsection 4.3.1 to improve the basic matrix factorization model. The process representation of matrix factorization exploiting topical context is illustrated in Figure 4.4.

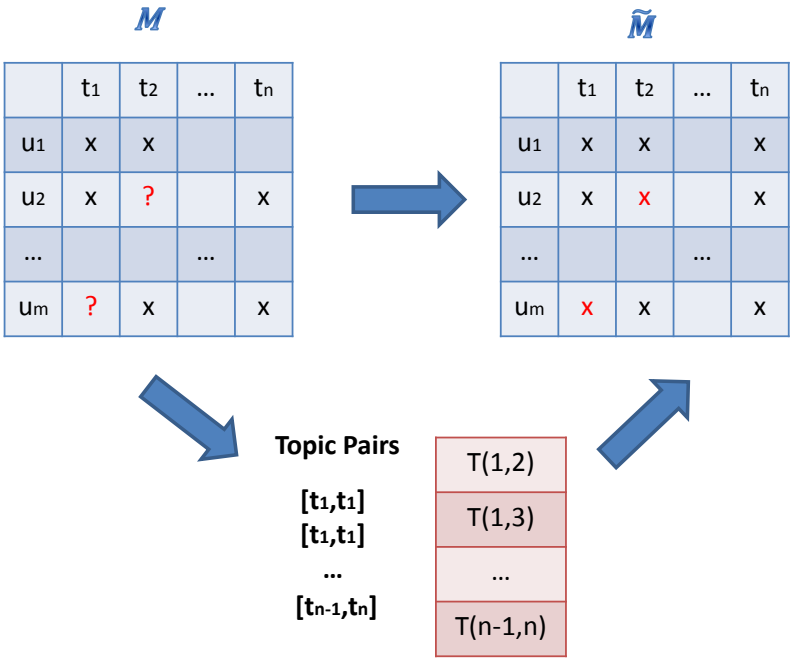


Figure 4.4: The process representation exploiting topical context.

We define the weight $T(i,j)$ between a pair of topics i and j according to the needs of different prediction tasks. The definitions will be presented in Chapter 5 and Chapter 6. Based on the general topical context hypothesis described in the above subsection 4.3.1, the topical context regularization are proposed to minimize the following terms:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} T(i,j) \|H(i,:) - H(j,:)\|_F^2 \tag{4.7}$$

where $T(i,j)$ is the weight value indicating how similar in content t_i and t_j are. The larger $T(i,j)$ is, the more similar two topics t_i and t_j are. If the value of $T(i,j)$ is small, the distance between two latent topic representations $H(i,:)$ and $H(j,:)$ can be large.

For a topic i, the terms in topical context regularization related to its latent

representation are,

$$\sum_{j=1}^{n} T(i,j)\|H(i,:) - H(j,:)\|_F^2 \tag{4.8}$$

In Eq. 4.8 we smooth the latent representation of i with other topics, and control these terms with weights between topic pairs, which makes we can get an approximate estimate even if topic i did not be discussed by many users in microblogging.

As the derivations in subsection 4.2.2, we can also get the matrix form of Eq. 4.7, thus

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} T(i,j)\|H(i,:) - H(j,:)\|_F^2 \\
&= \sum_{k=1}^{d} H^T(:,k)(D_T - T)H(:,k) \\
&= Tr(H^T L_T H).
\end{aligned} \tag{4.9}$$

Likewise, D_T is a diagonal matrix with the ith diagonal element $D_T(i,i) = \sum_{j=1}^{m} T(i,j)$, and $L_T = D_T - T$ is the Laplacian matrix. The matrix T including the weights between n topic pairs is as the following

$$T = \begin{bmatrix} 0 & T(1,2) & T(1,3) & \cdots & T(1,n) \\ T(2,1) & 0 & T(2,3) & \cdots & T(2,n) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T(n,1) & T(n,2) & T(n,3) & \cdots & 0 \end{bmatrix}$$

Hence, the model with topical context regularization can be formulated as:

$$\min_{U,H} \|Y \odot (M - UH^T)\|_F^2 + \lambda_1\|U\|_F^2 + \lambda_2\|H\|_F^2 + \beta Tr(H^T L_T H) \tag{4.10}$$

where $\beta \geq 0$ is the regularization parameter to control the regularization constraint

of topical context, balancing the reconstruction error between it and the front terms. Appropriate regularization parameter is also an important factor that leads to significant improvements for the prediction tasks. In Chapter 5 and Chapter 6, the regularization parameters will be determined through cross validation.

4.4 ScTcMF: The Proposed Framework with Social and Topical Context

In above sections, we formulated hypotheses about social context and topical context, and modeled regularization constraints with them, respectively. In this section, the Social context and Topical context incorporated Matrix Factorization (ScTcMF) framework is finally proposed.

Utilizing the social and topical context regularization constraints together, ScTcMF is formulated to *minimize* the following objective function:

$$\begin{aligned} F(U,H) &= \|Y \odot (M - UH^T)\|_F^2 + \lambda_1 \|U\|_F^2 + \lambda_2 \|H\|_F^2 \\ &\quad + \frac{\alpha}{2} \sum_{i=1}^{m} \sum_{j \in \mathcal{F}(i)} S(i,j) \|U(i,:) - U(j,:)\|_F^2 \\ &\quad + \frac{\beta}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} T(i,j) \|H(i,:) - H(j,:)\|_F^2 \\ &= \|Y \odot (M - UH^T)\|_F^2 + \lambda_1 \|U\|_F^2 + \lambda_2 \|H\|_F^2 \\ &\quad + \alpha Tr(U^T L_S U) + \beta Tr(H^T L_T H) \end{aligned} \tag{4.11}$$

where $\alpha, \beta \geq 0$ are respectively the social context regularization parameter and the topical context regularization parameter, and can be adjusted to make different impacts on the framework.

Note that when letting $\alpha = \beta = 0$, the ScTcMF degenerates to the basic matrix factorization. On the condition of $\alpha > 0, \beta = 0$ the framework only incorporates the

social context information; while on the condition of $\alpha = 0, \beta > 0$ the framework only incorporates the topical context information.

This objective function of Eq. 4.11 can be rewritten as

$$\begin{aligned} F(U,H) = Tr[&(Y^T \odot M^T)(Y \odot M) - (Y^T \odot M^T)(Y \odot UH^T) \\ &- (Y \odot M)(Y^T \odot HU^T) + (Y^T \odot HU^T)(Y \odot UH^T)] \\ &+ \lambda_1 Tr(U^T U) + \lambda_2 Tr(H^T H) \\ &+ \alpha Tr(U^T L_S U) + \beta Tr(H^T L_T H) \end{aligned} \tag{4.12}$$

Applying ScTcMF framework to the tasks of user interest prediction and user opinion prediction, both α and β are set to be positive for incorporating the information of social and topical context.

Chapter 5

User Interest Prediction

5.1 Problem Definition

The popular social networking service Twitter enriches and benefits people's daily life. At the same time, how to find out the really interesting and relevant topics from the massive streams of tweets, to provide precise topic recommendation for users, becomes a challenging problem in the real world. Previous collaborative filtering methods give solutions to traditional recommendation tasks considering users' positive reviews to help recommend items. However, for recommending interesting topics in microblogging, positive opinions toward a topic do not imply that user will be interested in it with high probability, for the user probably prefers to know those controversial topics or hot events with a large number of negative posts. In this chapter, we exploit the characteristics of topical opinion distribution to describe topical context information, and capture the weights between social friends under different opinion distribution topic patterns as social context information, for improving the performance of user interest prediction.

Given $\mathbf{u}$ be the set of m users, $\mathbf{t}$ be the set of n topics, in the task of user interest prediction, $I \in \mathbb{R}^{m \times n}$ is a user-topic interest matrix, with each element $I(u, i)$ representing the number of tweets tagged by user u on topic i. In this thesis,

we select the hashtags tagged by users as the topics. After information processing, the problem of predicting user interest can be reformulated as recommending the most possible topics that are interesting to users in microblogging for them.

5.2 Exploiting Social and Topical Context for Predicting User Interest

In this section, we first select a dataset for the task of user interest prediction from the real-world Twitter data we crawled. Subsection 5.2.2 describes the dataset used in this work, introduces the topical opinion distribution characteristics, and presents a series of observations on the dataset. Some findings about user interests in different topics are presented in subsection 5.2.3. Then, we describe how to formulate specific social context hypothesis and topical context hypothesis for user interest prediction. After validating the proposed hypotheses, we incorporate the information of social and topical context into ScTcMF framework according to them.

5.2.1 The Selected Dataset for User Interest Prediction

After the preprocessing presented in Chapter 3, we select a two month period real-world dataset (Nov 1 2010 - Dec 31 2010) from the crawled data for the task of user interest prediction. Then we computed a 5-core data, in which each user had interested in at least 5 different hashtags, and each hashtag was tagged by at least 5 different users. We list the statistics of the final dataset in Table 5.1.

Table 5.1: Statistics of the dataset for user interest prediction.

Statistics	Number
Users	4,306
Topics (Hashtags)	4,934
User-Topic Interests	155,021

5.2.2 Data Observations on Topical Opinion Distribution

In this subsection, we utilize the user opinions labeled by SentiStrength (with the method described in Chapter 3) to exploit the characteristics of topical opinion distribution. Based on the labeled user-topic opinions, we introduce the following characteristics to describe opinion distribution for each topic (hashtag).

- **Popularity:** This is actually the number of users who have been interested in the topic. It is utilized to estimate whether the topic is popular by a lot of users, which is also the total number of samples in a distribution.

- **Subjective Ratio:** This is the ratio of users who have obvious positive or negative opinions on the topic. Given a topic i, it is defined as:

$$Ratio(i) = \frac{N_p(i) + N_n(i)}{N(i)} \tag{5.1}$$

  where $N_p(i)$ and $N_n(i)$ represent the number of users giving positive and negative opinions on i, respectively. $N(i)$ is the number of all the users having tweeted on topic i. We assume that the topic interesting to users would arouse their subjective emotion more easily. A topic with high subjective ratio is more likely to interest a new user.

- **Opinion Entropy:** It measures how controversial the topic is to users, and is defined similarly to the entropy in Information theory:

$$Ent(i) = -(\frac{N_p(i)}{N_s(i)} \log \frac{N_p(i)}{N_s(i)} + \frac{N_n(i)}{N_s(i)} \log \frac{N_n(i)}{N_s(i)}) \tag{5.2}$$

  where $N_s(i)$ is the sum of $N_p(i)$ and $N_n(i)$. If the majority of users have positive or negative opinions on a topic, the value of opinion entropy will be low; if the users display a controversial debate on a topic, then the value of opinion entropy will be high. This characteristic may help find which type of topics is more interesting.

- **Average Positive Strength:** The opinion entropy only measures whether the topic is controversial to users, but not reflects the strength of user opinions. We calculate the average strength of positive opinions for each topic based on the strength scores provided by SentiStrength.

- **Average Negative Strength:** Similar to average positive strength, it is the average strength of negative opinions for each topic which is calculated based on the strength scores provided by SentiStrength.

We normalize all the values of these characteristics to $[0, 1]$ range, and study the correlations among them to see what can be discovered. From Figure 5.1 to Figure 5.3, we can observe that:

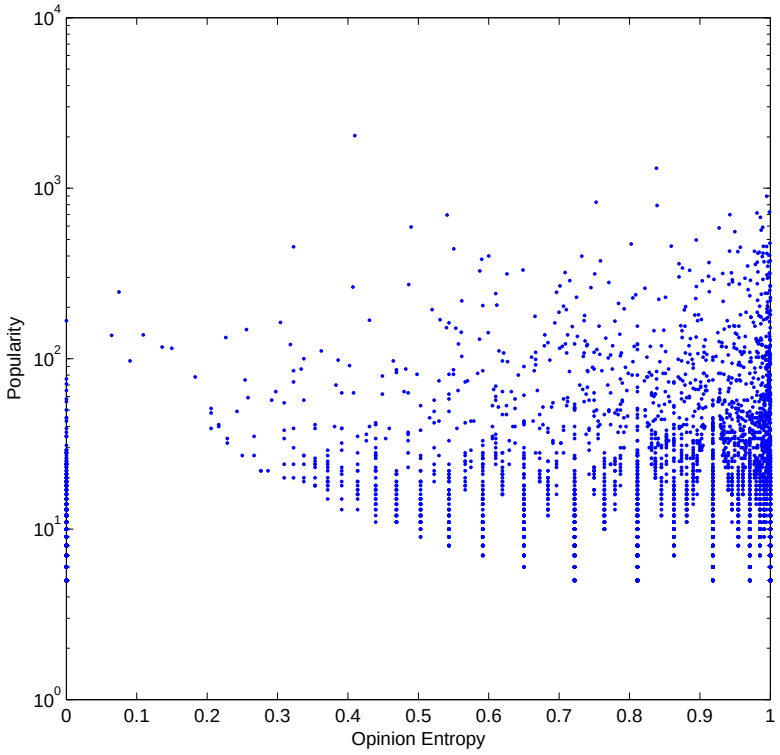


Figure 5.1: Popularity vs. Opinion Entropy

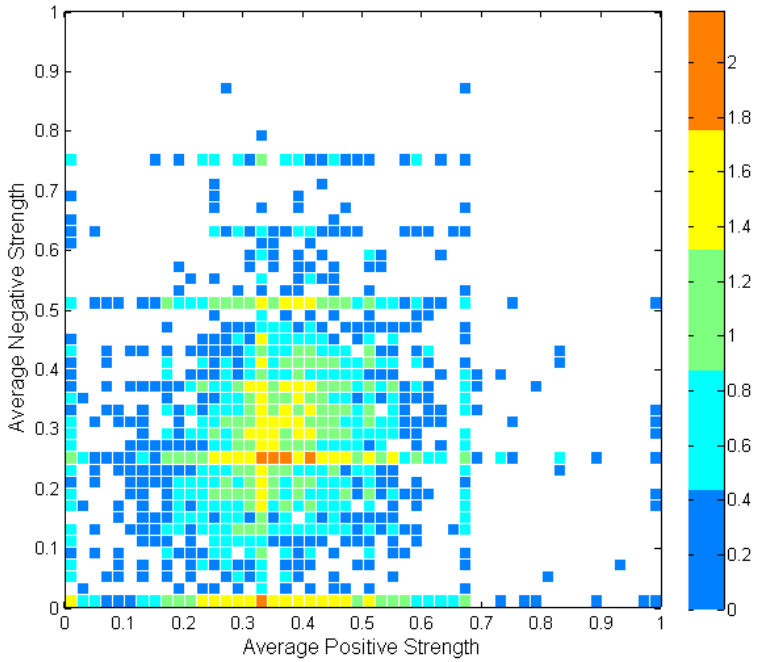


Figure 5.2: Ave Negative Strength vs. Ave Positive Strength

- The first observation is that topics tagged by more users tend to have higher opinion entropy, as shown in Figure 5.1. It also shows that most of topics tagged by more than 5 users are with relatively high opinion entropy.

- Figure 5.2 and Figure 5.3 are the heatmaps to plot topic distributions. The tick marks on the colorbars indicate log10 scale densities. In Figure 5.2, we observe that for most topics high average positive strength is along with high average negative strength, whereas there are also some exceptive topics with one high average strength value but the other average strength value is extremely low.

- We also average the opinion strength values of users no matter whether they are positive or negative, to get the average absolute opinion strength for each topic. The correlation between average absolute opinion strength and subjective ratio is shown in Figure 5.3. We find that the more users giving positive or negative opinions on a topic, the higher average absolute opinion strength it has from Figure 5.3.

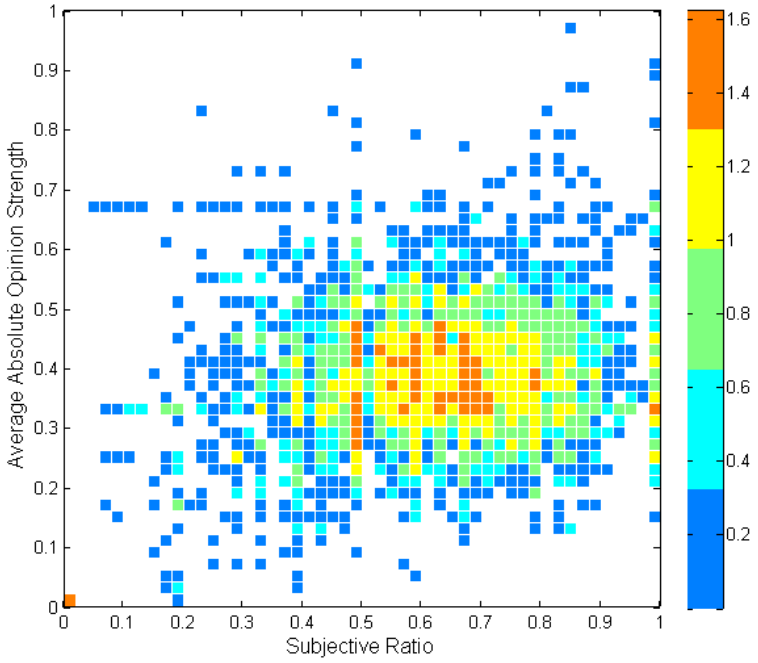


Figure 5.3: Average Absolute Opinion Strength vs. Subjective Ratio

5.2.3 Data Analysis on User Interests

We exploit the topical opinion distribution information in the last subsection. In related work, user information is also considered to be important and applied to the recommendation tasks [26, 99]. To further improve our approach, in this subsection, we divide the topics into several patterns based on their opinion distributions, and explore user interests under different patterns. We believe that some findings from these analyses will be helpful to utilize user information for the user interest prediction task. In detail, eight patterns are divided based on the three most important characteristics of opinion distribution: popularity, subjective ratio and opinion entropy. Each characteristic subspace is divided by the median value. All the eight topic patterns are listed in Table 5.2, in which each number stands for a unique topic pattern.

In Table 5.2, the letter L means the lower value group of the characteristic, while the letter H means the higher value group of the characteristic. The topics with all

Table 5.2: Topic patterns based on opinion distribution. (L=Low, H=High.)

Pattern	Popularity	Subjective Ratio	Opinion Entropy
1	L	L	L
2	L	L	H
3	L	H	H
4	L	H	L
5	H	L	L
6	H	L	H
7	H	H	H
8	H	H	L

three high characteristic values are usually about breaking events/news, celebrities, or Twitter memes. For example, *#iphone*,*#gossipgirl* and *#bieberfacts* are Pattern 7 topics. Pattern 2 topics only with high opinion entropy values are those less popular issues but deserving discussion, like *#backtothefuture*, *#poem*. Pattern 4 topics only with high subjective ratio values include *#vipfollow*, *#bestjonaslyrics*, and so on. Those Pattern 5 topics which are popular but with the other two low characteristic values, are also Twitter memes in most cases, such as *#peoplechoice*, and *#icantlivewithout*.

After defining topic patterns based on their opinion distribution characteristics, we investigate whether user interests under different patterns are significantly different. Let $Int_i(u)$ denote the average interest of user u in the topics of pattern i. Then we conduct a two-sample Kolmogorov-Smirnov test on the average user interest vectors Int_i and Int_j for each pair of patterns i and j. The null hypothesis is that the average user interests in Int_i and Int_j are from the same continuous distribution, and the alternative hypothesis is that they are from different continuous distributions.

We observe from Table 5.3 that for all the pairs of different patterns, the null hypothesis is rejected at the significant level 0.01. The p-values are very close to zero, which implies that the user interests under different topic patterns should be studied separately. We also note that the average user interest distributions of

Table 5.3: Statistics of user interest distribution difference.

p-value	2	3	4	5	6	7	8
1	$6.44e-81$	$4.60e-89$	$1.24e-89$	$<1e-200$	$<1e-200$	$<1e-200$	$<1e-200$
2	-	$3.29e-112$	$2.87e-68$	$<1e-200$	$<1e-200$	$<1e-200$	$<1e-200$
3	-	-	$9.19e-76$	$<1e-200$	$<1e-200$	$<1e-200$	$<1e-200$
4	-	-	-	$<1e-200$	$<1e-200$	$<1e-200$	$<1e-200$
5	-	-	-	-	$5.07e-33$	$1.63e-134$	$7.51e-57$
6	-	-	-	-	-	$1.31e-93$	$8.38e-87$
7	-	-	-	-	-	-	$1.16e-29$

topics with high popularity and those with low popularity are significantly different ($p < 1e-200$). This finding may be inspired to the work of detecting the trending topics in microblogging.

5.2.4 Incorporating Social Context

In recent work with social media data, researchers analyze the social network information to improve their recommendation tasks [62, 35]. In order to model social context regularization as introduced in Chapter 4 for the user interest prediction task in this dissertation, we investigate whether Twitter users with social friend relationships have more similar interests than those without in each different topic pattern, and propose the specific social context hypothesis in the task of user interest prediction as the following.

Hypothesis 3 *With high probability, the social friends hold more similar interests on the topics of different patterns than the non-friends.*

To validate this hypothesis, we conduct an analysis to show Twitter user interest similarities of social friends and those non-friends in each different topic pattern.

Under the topic pattern i, for every user u we calculate the mean of cosine similarities between u and his/her social friends, marked as $s_f(u,i)$; and the mean of similarities between u and randomly chosen users is marked as $s_r(u,i)$. The number of the randomly chosen users is set as the same as the number of u's social friends in the dataset. Then $\bar{s_f}(i)$ denotes the average mean of social friend similarities of

all users under pattern i, and $\bar{s_r}(i)$ denotes the average mean of random similarities of all users under pattern i. $\bar{v_f}(i)$ and $\bar{v_r}(i)$ are their variances respectively. Table 5.4 shows the results of $\bar{s_f}$, $\bar{s_r}$, $\bar{v_f}$ and $\bar{v_r}$ of each pattern over our dataset.

Table 5.4: The average means and variances of user interest similarities.

	Friends		Random	
Pattern	$\bar{s_f}$	$\bar{v_f}$	$\bar{s_r}$	$\bar{v_r}$
1	0.0071	0.0013	$9.17e-4$	$7.44e-5$
2	0.0062	$8.90e-4$	0.0011	$1.47e-4$
3	0.0059	$7.43e-4$	$8.51e-4$	$6.81e-5$
4	0.0086	0.0014	$7.49e-4$	$6.06e-5$
5	0.0343	0.0024	0.0152	$7.35e-4$
6	0.0347	0.0020	0.0188	$8.17e-4$
7	0.0781	0.0047	0.0417	0.0017
8	0.1468	0.0227	0.0873	0.0060

We find that in Table 5.4:

- The values of $\bar{s_f}$ is always larger than $\bar{s_r}$, which supports that Twitter users with social friendships have more similar interests than those without in each different topic pattern.

- The values of $\bar{s_f}$ under the last four topic patterns are much larger than those under the first four topic patterns, suggesting that the homophily between users and their social friends is more likely to happen when the number of the people taking behavior is larger.

- Meanwhile $\bar{v_f}$ is always larger than $\bar{v_r}$, which indicates that users adopt the topics interesting to their social friends differently. Thus they may have more similar interests with some friends than with others.

- The average mean similarities between users and the randomly chosen non-friends are small, so the average variances are small as well.

To capture user interest similarity between user i and j under different topic patterns, we utilize Jensen-Shannon Divergence [57, 64] to measure it, defined as

$$DJS(i,j) = \frac{1}{2}(DKL(p_i||m) + DKL(p_j||m)) \tag{5.3}$$

where DKL means the Kullback-Leibler Divergence, which can be calculated as

$$DKL(p_i||m) = \sum_k p_i(k) log \frac{p_i(k)}{m(k)} \tag{5.4}$$

In Eq. 5.3 and Eq. 5.4, $m = \frac{1}{2}(p_i + p_j)$, and p_i demotes the probability that user i is under pattern k [99], shown as follow:

$$p_i = \frac{n_i(k)}{n_i} \tag{5.5}$$

where n_i is the total number of topics interesting to user i, and $n_i(k)$ is the number of topics under pattern k that are interesting to user i.

At last, $S(i,j)$ in social context definition in the task of user interest prediction can be defined as

$$S(i,j) = \begin{cases} DJS(i,j) & \text{if } j \in \mathcal{F}(i) \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

5.2.5 Incorporating Topical Context

Social psychologists studied the distribution of opinion and observed it has influence on decision making of people [18]. In this task, we employ opinion distribution characteristics to help the topic recommendation for user. Therefore, we formulate the specific topical context hypothesis in user interest prediction as the following.

Hypothesis 4 *With high probability, two topics more similar in opinion distribution characteristics will interest users more similarly.*

In this subsection, we conduct a two-sample t-test to validate the above hypothesis. After normalizing all the values of these characteristics to $[0, 1]$ range, we calculate the cosine similarity ODS (Opinion Distribution Similarity) between opinion distribution vectors for topic pair i and j, as in Eq. 5.7.

$$ODS(i,j) = \frac{\sum_{k=1}^{K} Od(i,k) \cdot Od(j,k)}{\sqrt{\sum_{k=1}^{K} Od(i,k)^2}\sqrt{\sum_{k=1}^{K} Od(j,k)^2}} \tag{5.7}$$

where $Od(i,:)$ and $Od(j,:)$ denote the term frequency vectors of topic i and topic j respectively, and K is the number of features in the vectors.

Then we rank all pairs of topics according to their similarities in descending order, to form a higher-similarity group $\mathbf{h}$ in which are the top 10% topic pairs in term of their similarities, and a lower-similarity group $\mathbf{l}$ in which are the bottom 10% topic pairs. User interest similarities between topics are also calculated by cosine distance. Let $\mathbf{s_h}$ and $\mathbf{s_l}$ be user interest similarity vectors of topic pairs in $\mathbf{h}$ and $\mathbf{l}$ respectively. With these two vectors, we perform a two-sample t-test over the selected dataset for this task. The null hypothesis is $H_0 : \mathbf{s_h} \leq \mathbf{s_l}$, and the alternative hypothesis is $H_1 : \mathbf{s_h} > \mathbf{s_l}$. The null hypothesis is rejected at the significant level 0.01, which supports that with high probability two topics similar in opinion distribution characteristics will interest users similarly.

Consequently, $T(i,j)$ in topical context definition in the task of user interest prediction is defined as

$$T(i,j) = \begin{cases} ODS(i,j) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{5.8}$$

5.2.6 Details of ScTcMF Algorithm Solution

In this subsection, we introduce the detailed algorithm solution to solve the objective function proposed in Chapter 4.

In the task of user interest prediction in this chapter, the objective function F can be written as:

$$F(U,H) = \|Y \odot (I - UH^T)\|_F^2 + \lambda_1 \|U\|_F^2 + \lambda_2 \|H\|_F^2 + \alpha Tr(U^T L_S U) + \beta Tr(H^T L_T H \tag{5.9}$$

Because the user-topic interest matrix I is a non-negative matrix, we adopt an alternative optimization scheme proposed in previous work [20] to solve the objective function F. First, the derivatives of F with respect to U and H are:

$$\frac{\partial F}{\partial U} = -2(Y \odot I)H + 2Y \odot (UH^T)H + 2\lambda_1 U + 2\alpha L_S U \tag{5.10}$$

$$\frac{\partial F}{\partial H} = -2(Y \odot I)^T U + 2(Y \odot (UH^T))^T U + 2\lambda_2 H + 2\beta L_T H \tag{5.11}$$

Using the Karush-Kuhn-Tucker complementary condition,

$$[-(Y \odot I)H + Y \odot (UH^T)H + \lambda_1 U + \alpha L_S U](i,k)U(i,k) = 0 \quad \forall i \in [1,m], k \in [1,d] \tag{5.12}$$

$$[-(Y \odot I)^T U + (Y \odot (UH^T))^T U + \lambda_2 H + \beta L_T H](i,k)H(i,k) = 0 \quad \forall i \in [1,n], k \in [1,d] \tag{5.13}$$

which leads to the following updating formula of U and H according to the derivation process in [20].

$$U(i,k) \leftarrow U(i,k)\sqrt{\frac{[(Y \odot I)H + \alpha SU](i,k)}{[Y \odot (UH^T)H + \lambda_1 U + \alpha D_S U](i,k)}} \tag{5.14}$$

$$H(i,k) \leftarrow H(i,k)\sqrt{\frac{[(Y \odot I)^T U + \beta TH](i,k)}{[(Y \odot (UH^T))^T U + \lambda_2 H + \beta D_T H](i,k)}} \tag{5.15}$$

Apply the ScTcMF framework proposed in Chapter 4, the detailed algorithm solution for user interest prediction task is presented as below. We construct the matrices needed by the proposed algorithm at first, and then alternately update U and H until achieving convergence. Finally, we obtain a non-negative matrix $\tilde{I}$ that approximates the user-topic interest matrix.

Algorithm 1 ScTcMF Algorithm Solution for User Interest Prediction

Input: Social context matrix S, topical context matrix T, the observed user-topic interest matrix I, parameters $\lambda_1, \lambda_2, \alpha, \beta$

Output: The predicted user-topic interest matrix $\tilde{I}$

```
Initial U_0 randomly
Initial H_0 randomly
Construct the indicator matrix Y and the matrices D_S and D_T
while not convergent do
    for i=1 to m do
        for k=1 to d do
            Update U(i,k) ← U(i,k)√([(Y⊙I)H+αSU](i,k) / [Y⊙(UH^T)H+λ1U+αD_SU](i,k))
        end for
    end for
    for i=1 to n do
        for k=1 to d do
            Update H(i,k) ← H(i,k)√([(Y⊙I)^T U+βTH](i,k) / [(Y⊙(UH^T))^T U+λ2H+βD_TH](i,k))
        end for
    end for
end while
Compute Ĩ = UH^T
```

5.3 Experiments on User Interest Prediction

In this section, we present the experimental evaluations of our proposed ScTcMF framework on the task of user interest prediction. In subsection 5.3.1, we introduce the experiment setup, mainly including evaluation metrics and parameter settings. Next we compare the performance of different methods for user interest prediction on the selected Twitter dataset in subsection 5.3.2. In subsection 5.3.3, we discuss the computation cost issue about the implementation of the proposed ScTcMF framework. Finally, we analyze the effects of social and topical context regularization on improving the performance.

5.3.1 Experiment Setup

In this task, we computed a 5-core of the collected Twitter data, in which each user had interested in at least 5 different hashtags, and each hashtag was tagged by at least 5 different users. We list the statistics of the final dataset in Table 5.1.

We have selected the real-world Twitter dataset in subsection 5.2.1. In the experiments, for evaluating the performance of user interest prediction, we split the dataset into training data and test data by setting a timestamp (Dec 1 00:00:00 2010). Thus in the two-month period dataset, we use the data generated in the first month (Nov 1 2010 - Nov 30 2010) to train model, and apply it to the data generated in the second month (Dec 1 2010 - Dec 31 2010) for testing. All the observed Twitter data is organized into user-topic interest matrix I, whose each element $I(u, i)$ represents the number of tweets tagged by user u on topic i. After splitting, the training data includes 70,979 nonzero user-topic interest elements, while the test data includes 84,042 nonzero user-topic interest elements.

Note that the value of $I(u, i)$ ranges from 0 to a very large number. Instead of using the original values, we employ a mapping function $\frac{1}{1+x^{-1}}$ to bound the range of the values into $[0, 1]$, which should result in better performance as reported in

related work [22]. Besides, considering this task is a one-class collaborative filtering problem, in which zero elements in the matrix are either negative samples or missing data, we employ the sampling scheme presented in the work of Pan et al. [72] to get negative samples for training.

To measure the prediction quality in user interest prediction, we use the popular Top-N recommendation evaluation metrics $P@N$ and $R@N$ to evaluate the performance of precision and recall respectively [88]:

$$P@N = \frac{|\sum_{u \in \mathbf{u}} TopN(u) \bigcap T(u)|}{|\sum_{u \in \mathbf{u}} TopN(u)|} \tag{5.16}$$

$$R@N = \frac{|\sum_{u \in \mathbf{u}} TopN(u) \bigcap T(u)|}{|\sum_{u \in \mathbf{u}} T(u)|} \tag{5.17}$$

where $TopN(u)$ is the set of N topics recommended to user u that he/she has not tagged in the training data, and $T(u)$ is the set of topics tagged by user u in the test data. We set N to be 1, 5, 10 and 20 in our experiments.

The parameters applied in this task are determined through cross validation. For the proposed ScTcMF method, we choose $d = 5$ dimensions to represent the latent factor vectors. The values of λ_1 and λ_2 are set to be 0.01. The value of α is set to 0.05, while the value of β is set to 0.005.

5.3.2 Performance Comparison of User Interest Prediction

In the task of user interest prediction, we compare the proposed ScTcMF framework with several state-of-the-art methods, which are listed as follows.

- **Trending Topics (TT):** It sort all topics (hashtags) based on the number of tweets tagging them. This model actually recommends the most popular topics in the streams of tweets. Twitter recommended the overall trending topics to online users in this way in the early days. This naive baseline is

considered to be powerful because the crowds tend to heavily concentrate on a few of real-time tending topics in the sparse networking data [19].

- **Topic-Based Collaborative Filtering (TCF):** Item-based collaborative filtering is a state-of-the-art memory-based method for recommender systems. Since we aim to recommend the topics in Twitter, we employ the simple weighted average approach [96] to predict the user-topic interest, and mark it as TCF. Thus the interest of user u to topic i is predicted as

$$\tilde{I}(u,i) = \frac{\sum_{j \in T(u)} I(u,j) W(i,j)}{\sum_{j \in T(u)} W(i,j)} \tag{5.18}$$

  where $\tilde{I}(u,i)$ is the predicted value of interest. The summations are over all other topics tagged by user u. $W(i,j)$ is the weight between topic i and topic j. In this paper, we calculate the widely used cosine similarity between two vectors $I(:,i)$ and $I(:,j)$ in the training matrix as the weight between topic i and topic j.

- **Effective Missing Data Prediction (EMPD):** This is a memory-based collaborative filtering model proposed by Ma et al [61]. This method focuses on predicting the missing data with a combination of both user and item information. It is also able to determine whether to predict the missing data or not. Empirical studies have shown that the EMPD method is effective and more robust against data sparsity. In our experiment, the parameter λ balancing the information from users and items is tuned to 0.9 to achieve the best performance.

- **Non-Negative Matrix Factorization (NMF):** Non-negative matrix factorization [53] is also widely used in the collaborative filtering tasks. In our case, it infers non-negative user-topic interest by Eqs. 4.3, which is without neither social nor topical context regularization. The values of two smoothness

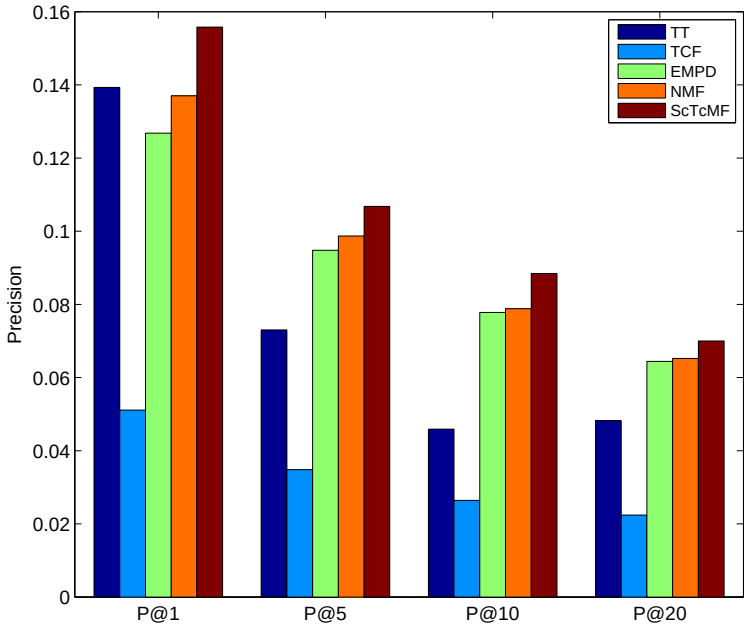


Figure 5.4: Precision performance of user interest prediction.

regularization parameters are set as the same as in ScTcMF framework.

The comparison results of precision and recall are shown in Figure 5.4 and Figure 5.5 respectively. From these results, we can obtain the following observations.

- The performances of the methods on the Twitter dataset are significantly different. ScTcMF performs the best in terms of both Top-N precision and recall, while TCF performs the worst. NMF consistently outperforms EMPD slightly. This observation indicates that the model-based methods are mostly superior to the memory-based methods on sparse data.

- The naive baseline method TT results in very good performance at P@1 and R@1, even better than the performance of EMPD and NMF. It is not surprisingly remembering the assumption that the crowds tend to heavily concentrate on a few of real-time tending topics in sparse networking data. Those most popular topics are usually interesting to new users. But TT loses its superiority as N grows. EMPD and NMF outperform it when N is no less than 5.

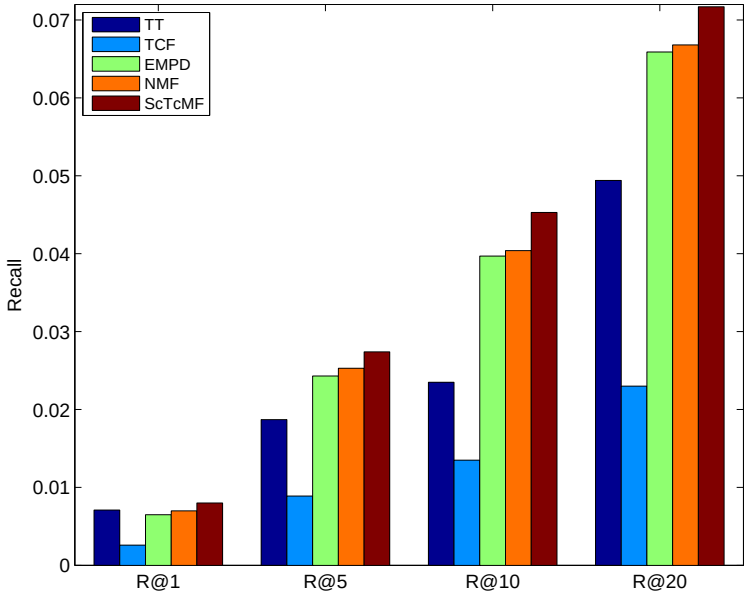


Figure 5.5: Recall performance of user interest prediction.

We are glad to see the proposed ScTcMF framework gives the best results all the time.

- ScTcMF always outperforms NMF, demonstrating the information of social and topical context does help improve the recommendation. Note that the ranges of the results of $P@N$ and $R@N$ vary as N grows from 1 to 20. We draw all the results with different N in the same figure, making some improvements seem slight, but they are significant indeed.

5.3.3 Time Complexity and Runtime Convergence

In this subsection, we discuss the computation cost issue about implementation. The time complexity of the proposed ScTcMF framework is $O(mnd)$. As introduced in Chapter 4, our proposed approach is based on the low-rank matrix factorization model, in which $d \ll min(m, n)$. The parameter d is often set to be a small value in the setting of implementation. The parameters m and n are determined by the size

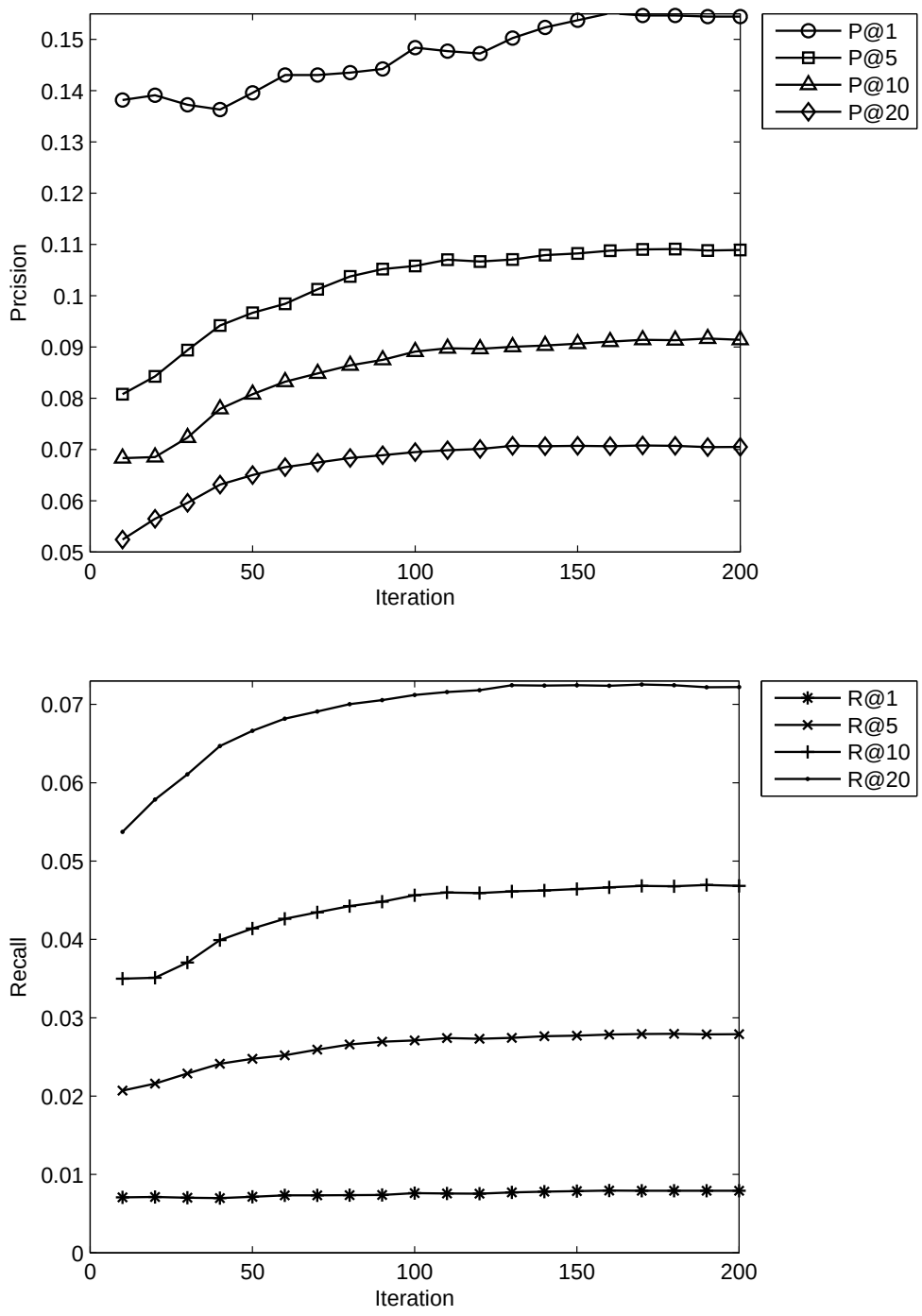


Figure 5.6: Runtime convergence of the ScTcMF method.

of dataset. During the implementation of ScTcMF, we analyze runtime convergence performance with different number of iterations.

Figure 5.6 shows the runtime convergence performance of ScTcMF on precision and recall. It is observed that both the precision and recall results converge after 200 iterations. This observation demonstrates that the implementation of the proposed approach is efficient and stable.

5.3.4 Effects of Social and Topical Context Regularization

In this task, we developed two regularization terms to incorporate social and topical context information. To further understand their effects on the performance of user interest prediction, we conduct experiments for analysis in this subsection.

At first, we investigate the effect of social context regularization. We compare the model only adding social context regularization with the NMF model. For each of them, we implement 5 times independently, and report the statistical results in Table 5.5. Similarly, in Table 5.6, we present the performance of the model only adding topical context regularization, to validate topical effect. Table 5.7 compares the statistical results of ScTcMF and NMF implementations. These three tables show the mean value with the standard deviation for all the precision and recall results, and indicate the percentage of improvement in the parentheses.

Table 5.5: The statistical effects of social context regularization.

	+ Social Context	NMF
$P@1$	$0.1437 \pm 0025(5.27\%)$	0.1365 ± 0.0019
$P@5$	$0.1011 \pm 0020(3.06\%)$	0.0981 ± 0.0010
$P@10$	$0.0843 \pm 6.89e-4(7.12\%)$	$0.0787 \pm 6.50e-4$
$P@20$	$0.0675 \pm 6.16e-4(3.21\%)$	$0.0654 \pm 5.68e-4$
$R@1$	$0.0074 \pm 1.26e-5(6.76\%)$	$0.0069 \pm 5.48e-5$
$R@5$	$0.0259 \pm 5.50e-4(2.78\%)$	$0.0252 \pm 1.14e-4$
$R@10$	$0.0432 \pm 3.56e-4(6.67\%)$	$0.0405 \pm 2.88e-4$
$R@20$	$0.0692 \pm 6.66e-4(3.59\%)$	0.0668 ± 0.0011

Table 5.6: The statistical effects of topical context regularization.

	+ Topical Context	NMF
$P@1$	$0.1485 \pm 8.70e-4(8.79\%)$	0.1365 ± 0.0019
$P@5$	$0.1038 \pm 9.44e-4(5.81\%)$	0.0981 ± 0.0010
$P@10$	$0.0837 \pm 0.0017(6.35\%)$	$0.0787 \pm 6.50e-4$
$P@20$	$0.0675 \pm 7.83e-4(3.21\%)$	$0.0654 \pm 5.68e-4$
$R@1$	$0.0076 \pm 7.07e-5(10.14\%)$	$0.0069 \pm 5.48e-5$
$R@5$	$0.0266 \pm 2.88e-4(5.55\%)$	$0.0252 \pm 1.14e-4$
$R@10$	$0.0434 \pm 8.23e-4(7.16\%)$	$0.0405 \pm 2.88e-4$
$R@20$	$0.0697 \pm 0.0010(4.34\%)$	0.0668 ± 0.0011

Table 5.7: The statistical results of ScTcMF vs. NMF

	ScTcMF	NMF
$P@1$	$0.1553 \pm 7.19e-4(13.77\%)$	0.1365 ± 0.0019
$P@5$	$0.1071 \pm 0.0015(9.17\%)$	0.0981 ± 0.0010
$P@10$	$0.0892 \pm 0.0013(13.34\%)$	$0.0787 \pm 6.50e-4$
$P@20$	$0.0698 \pm 9.52e-4(6.73\%)$	$0.0654 \pm 5.68e-4$
$R@1$	$0.0080 \pm 3.65e-5(15.94\%)$	$0.0069 \pm 5.48e-5$
$R@5$	$0.0274 \pm 3.85e-4(8.73\%)$	$0.0252 \pm 1.14e-4$
$R@10$	$0.0457 \pm 6.75e-4(12.84\%)$	$0.0405 \pm 2.88e-4$
$R@20$	$0.0715 \pm 9.88e-4(7.04\%)$	0.0668 ± 0.0011

From Table 5.5 to Table 5.7, we observe the following:

- Both the model only adding social context regularization and the model only adding topical context regularization can improve over the NMF baseline, which validates the effects of social context regularization and topical context regularization respectively.

- By incorporating both social and topical context regularization, ScTcMF outperforms NMF significantly, suggesting the proposed ScTcMF framework that captures different types of context information for the user interest prediction task is successful.

- The improvements of ScTcMF over NMF are most significant at $P@1$ and $R@1$, indicating that the proposed ScTcMF framework helps to finding the most interesting topics for microblogging users. The improvements of ScTcMF are smaller when the setting of N is larger than 10.

Chapter 6

User Opinion Prediction

6.1 Problem Definition

With more and more people sharing their opinions freely using microblogging, sentiment analysis and opinion mining on the text content of Twitter has been extensively studied by researchers in recent years. Most of the early work applied the state-of-the-art methods of sentiment analysis and opinion mining to detect tweet sentiment. This part of work can be regarded as tweet-level research. Some work proposed approaches to mining user-level opinions later. However, several fixed topics were chosen for opinion mining, and the proposed approaches did not take the correlations among topics into account. Our task of user opinion prediction attempts to solve a new problem different from the existing work. We focus on predicting user-topic level opinions before observing the corresponding content of tweets, which is a novel problem providing both challenges and opportunities for research [83]. Figure 6.1 illustrates the task of user-topic level opinion prediction, which is referred to as user opinion prediction for short in this dissertation.

In the task of user opinion prediction, given $\mathbf{u}$ be the set of m users, $\mathbf{t}$ be the set of n topics, a user-topic matrix $O \in \mathbb{R}^{m \times n}$ denotes the opinion label matrix, consisting of elements $O(u, i)$, which represents the opinion of user u for topic i.

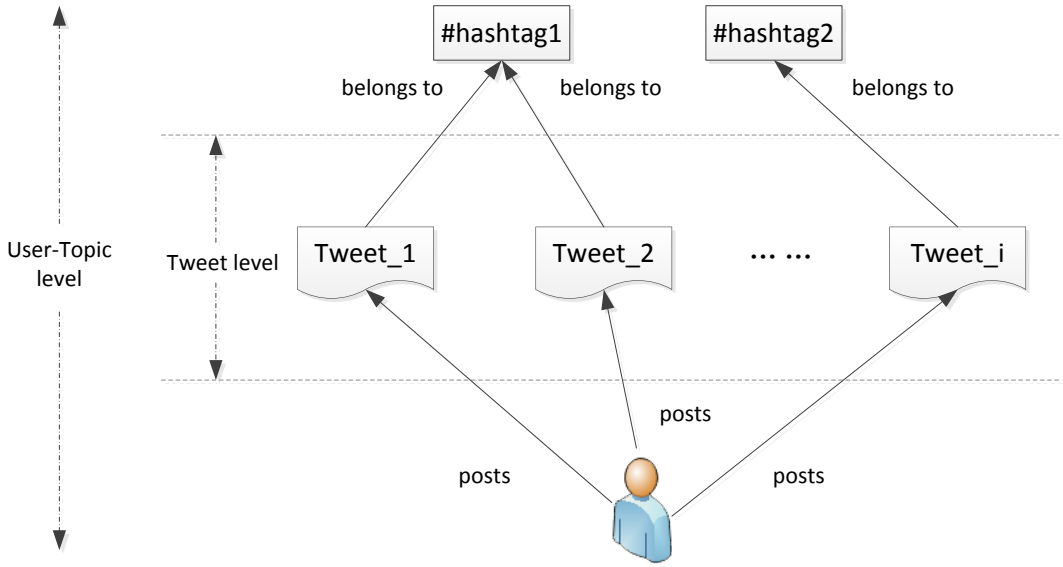


Figure 6.1: User-topic level opinion prediction.

Like the classification used in most of sentiment analysis and opinion mining tasks, we simply define $O(u, i) = 1$ as the positive opinion label, and $O(u, i) = -1$ as the negative one. The neutral ones are not considered in this task. If there is no observed opinion label user u gave to topic i, the element $O(u, i)$ will be assigned 0. The problem we want to study is then turned into how to predict the missing opinion labels in the user-topic opinion matrix O by employing the observed data from Twitter.

6.2 Exploiting Social and Topical Context for Predicting User Opinion

In this section, we describe specific social context hypothesis and topical context hypothesis for user opinion prediction, and present how to incorporate them into

ScTcMF framework respectively.

6.2.1 Incorporating Social Context

Due to the mechanisms of Twitter we have introduced in Chapter 4, we consider that the flow of information between the followers and friends on Twitter is bidirectional, and both ends of a following relationship will more or less influence the opinions of each other, via their expression in tweets. For obtaining the mutual opinion influence on Twitter, we define both a user's friends and followers as his/her social friends, and convert the directed following relationship network into an undirected social friend relationship network. Subsequently, the social context hypothesis in user opinion prediction task can be described as the following.

Hypothesis 5 *With high probability, the social friends hold more similar opinions on the topics than the non-friends.*

As mentioned in Chapter 4, the value of $S(i, j)$ indicates the weight between social friends i and j. In this task, we directly calculate the cosine similarity between the two corresponding row vectors $O(i, :)$ and $O(j, :)$ of the user-topic opinion matrix O, to capture the difference of social friends' opinions towards different topics, and define it as UOS (User Opinion Similarity), thus

$$UOS(i,j) = \frac{\sum_{k=1}^{n} O(i,k) \cdot O(j,k)}{\sqrt{\sum_{k=1}^{n} O(i,k)^2} \sqrt{\sum_{k=1}^{n} O(j,k)^2}}. \tag{6.1}$$

According to the definition of the weight values in S, we apply a mapping $UOS(i, j) = (UOS(i, j) + 1)/2$ to bound the range of UOS similarity into $[0, 1]$. Then the element $S(i, j)$ can be formally defined as:

$$S(i,j) = \begin{cases} UOS(i,j) & \text{if } j \in \mathcal{F}(i) \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

6.2.2 Incorporating Topical Context

In previous research tasks of information retrieval and text mining, content-based topic correlations are studied to improve the tasks [8, 113, 37]. Inspired by those researches, we also exploit the content-based correlations between topics for predicting the unknown opinions in the user-topic opinion matrix. The hypothesis we model for topical context is as the following.

Hypothesis 6 *With high probability, two topics more similar in content will be given more similar opinions by the users.*

In the work of [109], the authors proposed different similarity measures using the topic distributions and association. We make a comparison on several measures they mentioned, and finally choose the cosine similarity for its simplicity and efficiency. Taking unique terms appeared in the tweets collection (after stop words removal) as features, and term frequency as the feature value, term frequency vector $tf(i)$ could be created for each topic i, and the cosine similarities between term frequency vectors could be calculated to measure the content-based similarities between the corresponding topics, which we mark as TCS (Topic Content Similarity).

$$TCS(i,j) = \frac{\sum_{k=1}^{N} tf(i,k) \cdot tf(j,k)}{\sqrt{\sum_{k=1}^{N} tf(i,k)^2}\sqrt{\sum_{k=1}^{N} tf(j,k)^2}} \tag{6.3}$$

where $tf(i,:)$ and $tf(j,:)$ denote the term frequency vectors of topic i and topic j respectively, and N is the number of features in the vectors. In this definition, the similarity values range from 0 to 1, since the term frequencies cannot be negative. Finally the element $T(i,j)$ can be presented as follows

$$T(i,j) = \begin{cases} TCS(i,j) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

6.2.3 Details of ScTcMF Algorithm Solution

Note that the user-topic opinion matrix O is not a non-negative matrix. In this case, the gradient based approaches are simple and effective among the existing optimization techniques. In this dissertation, we apply a standard gradient descent method to solve the objective function in Eq. 6.5. In the method, U_{t+1} and H_{t+1} are updated in each step as:

$$U_{t+1} \leftarrow U_t - \gamma \frac{\partial F_t}{\partial U_t}) \tag{6.5}$$

$$H_{t+1} \leftarrow H_t - \gamma \frac{\partial F_t}{\partial H_t}) \tag{6.6}$$

In the above equations, γ is the step size to make control. $\frac{\partial F_t}{\partial U_t}$ and $\frac{\partial F_t}{\partial H_t}$ are the partial derivatives to U and H respectively, which are employed as the gradients in the $t+1$ step.

Apply the ScTcMF framework proposed in Chapter 4, the detailed algorithm solution for user opinion prediction task is shown in Algorithm 2. From line 1 to line 3, we initial the matrices needed by the algorithm. From line 4 to line 9, we update U and H along the negative gradient direction until achieving convergence. In the end, we obtain a matrix $\tilde{O}$ including the predicted opinions.

6.3 Experiments on User Opinion Prediction

In this section, we present the experimental evaluation of our proposed ScTcMF framework with detailed discussions. We begin by introducing the experiment setup. Then we test the hypotheses proposed for user opinion prediction in this chapter on the selected dataset. Next, we evaluate and compare the performance of different

Algorithm 2 ScTcMF Algorithm Solution for User Opinion Prediction

Input: Social context matrix S, topical context matrix T, the set of labeled user-topic opinions O_0, parameters $\lambda_1, \lambda_2, \alpha, \beta$, step size γ
Output: The predicted user-topic opinion matrix $\tilde{O}$

Initial U_0 randomly
Initial H_0 randomly
Construct the indicator matrix Y and the Laplacian matrices L_S and L_T
while not convergent **do**
    Compute $\frac{\partial F_t}{\partial U_t}$)
    Compute $\frac{\partial F_t}{\partial H_t}$)
    Set $U_{t+1} \leftarrow U_t - \gamma \frac{\partial F_t}{\partial U_t}$)
    Set $H_{t+1} \leftarrow H_t - \gamma \frac{\partial F_t}{\partial H_t}$)
end while
Set $U = U_{t+1}$
Set $H = H_{t+1}$
Compute $\tilde{O} = UH^T$

methods for user-topic opinion prediction. Lastly, we investigate the impact of the regularization parameters.

6.3.1 Experiment Setup

In this task, we select popular hashtags in real-world data as hot topics. Therefore, firstly we gathered the hashtags those with a frequency more than 100. Secondly, considering the practicality of our solution, we only chose those debated topics from the high-frequency hashtags. Let C_h^+ denote the count of people whose opinions towards hashtag h are positive, and C_h^- denote the count of people whose opinions towards h are negative. We only kept the hashtags which satisfied the following condition in Eq. 6.7,

$$\theta_1 < C_h^+ / C_h^- < \theta_2 \tag{6.7}$$

where θ_1 and θ_2 are threshold values, used for excluding the topics hugely biased towards positive or negative. We set θ_1 to 0.2 and θ_2 to 5 in this work. Besides, we kept the hashtag tagged by at least 5 different users. Finally, we obtained $1,335$ hashtags meeting all the conditions as the hot topics in our task.

The statistics of the final dataset for user opinion prediction experiments are

listed in Table 6.1.

Table 6.1: Statistics of the dataset for user opinion prediction.

Statistics	Number
Users	3,485
Topics (Hashtags)	1,335
User-Topic Opinions	102,569

For evaluating the proposed method via experiments, we randomly split the selected dataset into training and testing sets. To study the impact of different training data sets on the performance, we respectively select 10%, 20%, 50%, 80%, and 90% of the whole opinion labels randomly as the training data, to predict the remaining portions of opinion labels. The random selection was carried out 5 times independently, and we then report average results.

Two popular prediction metrics, Root Mean Square Error (RMSE) and Accuracy, are used to measure the prediction quality in our task. The metric of RMSE is defined as:

$$RMSE = \sqrt{\frac{1}{N_T}\sum_{i,j}(O(u,i) - \tilde{O}(u,i))^2} \tag{6.8}$$

where N_T denotes the number of opinions for testing. As mentioned in Section 2, we let $O(u,i) = 1$ to denote positive user-topic opinion, and $O(u,i) = -1$ to denote negative user-topic opinion from user u to topic i. However, the value of $\tilde{O}(u,i)$ obtained by the prediction method may be not an integer which exactly equals to 1 or -1. Hence, before calculating the Accuracy, we map the value of $\tilde{O}(u,i)$ with the *sign function*. Thus $\tilde{O}(u,i)$ will be mapped to 1 if it is positive, and -1 if negative. Then we conduct Accuracy calculation with the mapped values. In the experiments, a smaller RMSE value or a higher Accuracy value indicates a better prediction performance.

6.3.2 Hypotheses Testing

Before going further to evaluate the prediction performance, we validate the hypotheses about the social context and topical context proposed in subsection 6.2.1 and subsection 6.2.2 over the selected dataset.

Social Context hypothesis testing: For testing the hypothesis about social context that we proposed in Section 3.2, we present hypothesis testing to validate the opinion homophily between users and their friends, and the opinion homophily between users and their followers respectively. In the first testing, for each user u, we calculate average UOS between u and his/her friends, marked as $s_{fr}(u)$; and UOS between this user and randomly chosen users, marked as $s_{rr}(u)$. The number of the randomly chosen users is set to be the same as the number of u's friends in the dataset. Finally we obtain two vectors $\mathbf{s_{fr}}$ and $\mathbf{s_{rr}}$, and then conduct a two-sample t-test on them. The null hypothesis is $H_0 : \mathbf{s_{fr}} \leq \mathbf{s_{rr}}$, and the alternative hypothesis is $H_1 : \mathbf{s_{fr}} > \mathbf{s_{rr}}$. In our dataset, the null hypothesis is rejected at the 0.0005 significant level with p-value of $2.7e$-37. Next, for each user u, we calculate average UOS between u and his/her followers, marked as $s_{fo}(u)$; and UOS between this user and randomly chosen users, marked as $s_{ro}(u)$. The number of the randomly chosen users is also set to be the same as the number of u's followers in the dataset. Then we obtain two vectors $\mathbf{s_{fo}}$ and $\mathbf{s_{ro}}$, and then conduct a two-sample t-test on them. The null hypothesis is $H_0 : \mathbf{s_{fo}} \leq \mathbf{s_{ro}}$, and the alternative hypothesis is $H_1 : \mathbf{s_{fo}} > \mathbf{s_{ro}}$. The null hypothesis is rejected at the 0.0005 significant level with p-value of $9.2e$-18. The hypothesis testing results indicate that the opinion homophily exists among users and his/her social friends (both friends and followers), and thus social friend relationships could be exploited in user opinion prediction.

Topical Context hypothesis testing: For testing the hypothesis about topical context that we proposed in Section 3.3, we let $\mathbf{h}$ be the set of the topic pairs (i, j) with the 10% highest $TCS(i, j)$, and $\mathbf{l}$ be the set of the topic pairs (i, k) with the 10% lowest $TCS(i, k)$. We then calculate the prior opinion similarities for these

topic pairs, which we defined as TOS (Topic Opinion Similarity):

$$TOS(i,j) = \frac{\sum_{u=1}^{m} O(u,i) \cdot O(u,j)}{\sqrt{\sum_{u=1}^{m} O(u,i)^2}\sqrt{\sum_{u=1}^{m} O(u,j)^2}} \tag{6.9}$$

where $O(:,i)$ and $O(:,j)$ are the corresponding column vectors of i and j in the user-topic opinion matrix O. Using the same mapping function we applied to UOS in Section 3, we map the range of this similarity into $[0,1]$. Then we mark $\mathbf{s_h}$ and $\mathbf{s_l}$ as the vectors of TOS values between topic pairs in $\mathbf{h}$ and $\mathbf{l}$ respectively. Subsequently, we validate this hypothesis over our dataset by using a two-sample t-test. The null hypothesis is $H_0 : \mathbf{s_h} \leq \mathbf{s_l}$, and the alternative hypothesis is $H_1 : \mathbf{s_h} > \mathbf{s_l}$. The null hypothesis is rejected at the 0.0005 significant level with p-value of $1.91e$-44. The evidence from this t-test supports that with higher probability, the users hold consistent opinions on the topics with similar content.

6.3.3 Performance Comparison of User Opinion Prediction

As the task of user opinion prediction is modeled as a new collaborative filtering problem, we compare the proposed ScTcMF framework with the state-of-the-art methods in collaborative filtering. The models only with social context or only with topical context are also used for comparison. All of these baseline methods are listed as the following.

- *TopicMean*: with this method, the opinion which a user gives to a topic is predicted by the mean value of known opinions the user gave.
- *UCF*: the memory-based approaches are the most popular prediction methods, and are widely adopted in commercial collaborative filtering systems [96]. UCF (User-based Collaborative Filtering) is a typical memory-based approach, by which the opinion a user u gives to a topic i is calculated as an aggregation of

the similar users' opinions towards the topic:

$$O(\tilde{u}, i) = O(\bar{u}, :) + \frac{\sum_{u,v \in \mathbf{u}} UOS(u, v) \cdot (O(v, i) - O(\bar{v}, :))}{\sum_{u,v \in \mathbf{u}} UOS(u, v)} \tag{6.10}$$

where $O(\bar{u}, :)$ and $O(\bar{v}, :)$ are the average values of the opinions users u and v gave, and UOS is utilized to measure the similarity between u and v.

- *MF*: the basic low-rank matrix factorization model, which is as shown in Eq. 4.3. The basic MF model is also widely used in the traditional collaborative filtering tasks.

- *ScMF*: this method employs the model only with social context regularization constraint we formulated in Eq. 4.6.

- *SfMF*: it is also a matrix factorization based method incorporating the social network information. But in this method we discard the similarity calculation by setting all the weight values between social friends to 1. We design this method to examine if using UOS as the weight value in S contributes to the regularization constraint of social context.

- *TcMF*: this method employs the model only with topical context regularization constraint we formulated in Eq. 4.10.

- *ToMF*: it is also a matrix factorization based method incorporating the topic correlation information. Different from TcMF, the underlying hypothesis of ToMF is *with high probability, two topics have been given similar opinions by the users before will be given more similar opinions in the future.* With the same regularization term that we proposed in Eq. 4.10, in this method, the elements in the adjacency matrix T are calculated as follows,

$$T(i, j) = \begin{cases} TOS(i, j) & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

This topic-oriented hypothesis has been utilized in some other collaborative filtering tasks [26, 61]. We compare ToMF with TcMF to examine that using which hypothesis can model topical context better for user-topic opinion prediction.

In the experiments, the values of λ_1 and λ_2 in all the low-rank matrix factorization based methods are set to 1, and the latent feature dimension d is set to 5 based on the results of pre-performed testing for parameter tuning. In all the methods using regularization constraints, we adjust the regularization parameters for them and present their best performance. In the proposed ScTcMF method, the regularization parameters α and β are tuned to 10 and 0.01 respectively. In ScMF, the value of α is set to 10; in SfMF, the value of α is set to 1; in TcMF and ToMF the value of β is set to 0.01 to achieve their best performance. The experimental results measured by RMSE and Accuracy are shown in Table 6.2 and Fig. 6.2 respectively.

Table 6.2: RMSE comparisons using different training sets.

Training set	*TopicMean*	*UCF*	*MF*	*ScMF*	*SfMF*	*TcMF*	*ToMF*	*ScTcMF*
10%	1.2535	1.1153	1.0117	0.9949	0.9973	0.9853	0.9853	**0.9771**
20%	1.2948	1.0521	0.9938	0.9747	0.9773	0.9715	0.9729	**0.9645**
50%	1.2417	0.9923	0.9826	0.9609	0.9648	0.9643	0.9654	**0.9561**
80%	1.1433	0.9731	0.9758	0.9549	0.9597	0.9600	0.9612	**0.9513**
90%	1.1109	0.9687	0.9727	0.9512	0.9563	0.9575	0.9583	**0.9488**

In our experiments, the values of λ_1 and λ_2 in all the low-rank matrix factorization based methods are set to 1, and the latent feature dimension d is set to 5 based on the results of pre-performed testing for parameter tuning. In all the methods using regularization constraints, we adjust the regularization parameters for them and present their best performance. In the proposed ScTcMF method, the regularization parameters α and β are tuned to 10 and 0.01 respectively. In ScMF, the value of α is set to 10; in SfMF, the value of α is set to 1; in TcMF and ToMF the value of β is set to 0.01 to achieve their best performance. More details about the impact of α and β will be discussed in the next subsection. The experimental

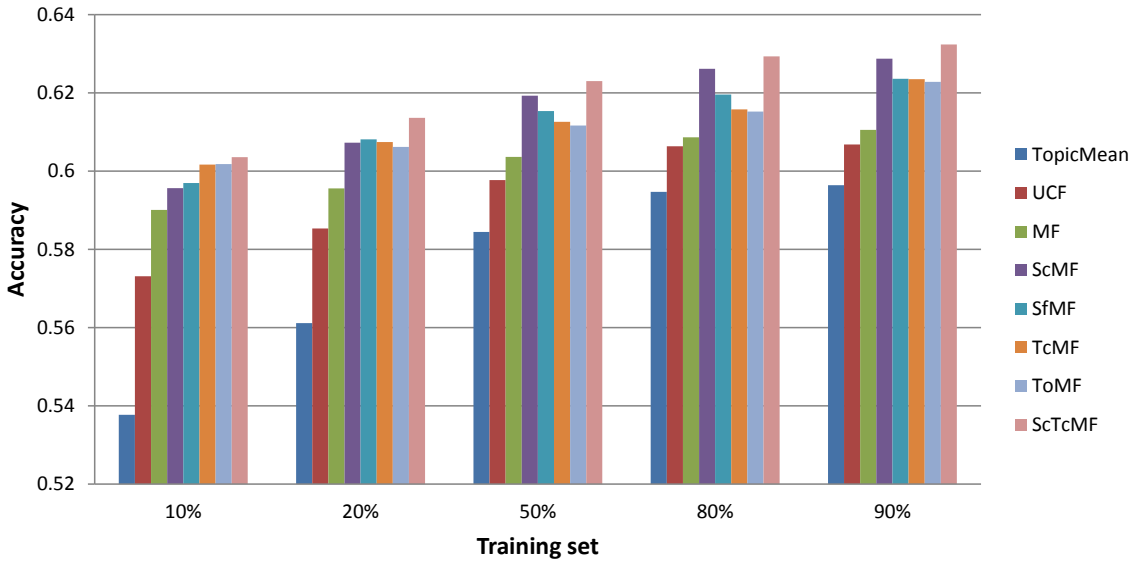


Figure 6.2: Accuracy comparisons using different training sets.

results measured by RMSE and Accuracy are shown in Table 6.2 and Figure 6.2 respectively. In Table 6.2, the best result of each line is bold. From these results, we can obtain the following observations.

- All the results of Accuracy in Fig. 6.2 are better than that of randomly guessing (which is 0.5 in our task). The matrix factorization based methods with regularization constraints consistently outperform the TopicMean and UCF methods, and most of them can improve the basic MF model significantly, both by Accuracy and RMSE. The methods with social context get more improvement than those with topical context except when the training data is extremely sparse. ScTcMF always generates the best results.

- The matrix factorization based methods can work well even when the training data is very sparse. The smaller the size of training data, the more performance improvement can be achieved. All the matrix factorization models incorporating regularization constraints generate better results than MF, indicating that the regularization constraints actually benefit the user opinion prediction.

- Comparing SfMF with ScMF, we can find that ScMF always performs better than SfMF. This observation demonstrates that the opinion homophily exists between social friend pairs, but not all the social friends hold extremely similar opinions on the hot topics, so an effective function to calculate the similarities between social friends is important.

- As to TcMF and ToMF, TcMF can generate slightly better results than TOMF, but the improvement is stable, which indicates the content-based correlations among topics are more useful for modeling topical context in the user opinion prediction task.

- Using the social and topical context regularization constraints together, ScTcMF obtains better results than using them respectively.

6.3.4 Analysis and Discussion on User Opinion Prediction

In the previous subsection, we used RMSE and accuracy to measure the overall prediction quality of each method, but did not analyze different kinds of errors. As mentioned before, there are only two types of opinion labels in our test set: positive and negative. For formulation, we let 1 denote positive opinion label, and -1 denote negative opinion label. The predicted values by all of our methods are real numbers, and we map these values with the sign function to make positive/negative classification. Note that in our experiments, all the predicted values do not equal the threshold 0, which means there is no error caused by missing value. In this section, we present the method performance measured by *precision*, *recall* and *F1 score* on positive class and negative class respectively, in order to give deeper insight into the incorrect prediction. To save space, we only show the results of the experiments using 10%, 50% and 90% training data.

The results from Table 6.3 to Table 6.5 (the best average results are bold) show that all the methods always achieve better performances on positive opinion class,

Table 6.3: Precision comparisons in positive and negative opinion prediction

	Metrics	10%	50%	90%
Positive	*TopicMean*	0.6238	0.6287	0.6362
	UCF	0.6310	0.6320	0.6393
	MF	0.6228	0.6487	0.6617
	ScMF	0.6332	0.6621	0.6769
	SfMF	0.6309	0.6591	0.6720
	TcMF	0.6111	0.6427	0.6623
	ToMF	0.6135	0.6426	0.6627
	ScTcMF	0.6157	0.6569	0.6737
Negative	*TopicMean*	0.4449	0.4891	0.4932
	UCF	0.4765	0.5112	0.5129
	MF	0.4988	0.5179	0.5148
	ScMF	0.5082	0.5399	0.5405
	SfMF	0.5110	0.5342	0.5335
	TcMF	0.5407	0.5373	0.5375
	ToMF	0.5363	0.5352	0.5359
	ScTcMF	o.5400	0.5500	0.5487
Average	*TopicMean*	0.5343	0.5589	0.5647
	UCF	0.5538	0.5716	0.5761
	MF	0.5607	0.5833	0.5883
	ScMF	0.5707	0.6010	0.6087
	SfMF	0.5710	0.5967	0.6028
	TcMF	0.5759	0.5900	0.5999
	ToMF	0.5749	0.5889	0.5993
	ScTcMF	**0.5779**	**0.6035**	**0.6110**

Table 6.4: Recall comparisons in positive and negative opinion prediction

	Metrics	10%	50%	90%
Positive	*TopicMean*	0.5479	0.7259	0.7649
	UCF	0.6681	0.7651	0.7915
	MF	0.7761	0.7194	0.7183
	ScMF	0.7500	0.7275	0.7299
	SfMF	0.7656	0.7241	0.7285
	TcMF	0.8956	0.7767	0.7605
	ToMF	0.8807	0.7735	0.7567
	ScTcMF	0.8751	0.7589	0.7515
Negative	*TopicMean*	0.5231	0.3797	0.3435
	UCF	0.4360	0.3554	0.3296
	MF	0.3217	0.4362	0.4488
	ScMF	0.3729	0.4627	0.4768
	SfMF	0.3535	0.4580	0.4661
	TcMF	0.1775	0.3751	0.4178
	ToMF	0.1991	0.3775	0.4217
	ScTcMF	0.2116	0.4264	0.4534
Average	*TopicMean*	0.5355	0.5528	0.5592
	UCF	0.5521	0.5603	0.5606
	MF	0.5489	0.5778	0.5836
	ScMF	**0.5615**	**0.5951**	**0.6034**
	SfMF	0.5596	0.5928	0.5973
	TcMF	0.5366	0.5762	0.5892
	ToMF	0.5399	0.5755	0.5892
	ScTcMF	0.5434	0.5927	0.6024

Table 6.5: F1-Score comparisons in positive and negative opinion prediction

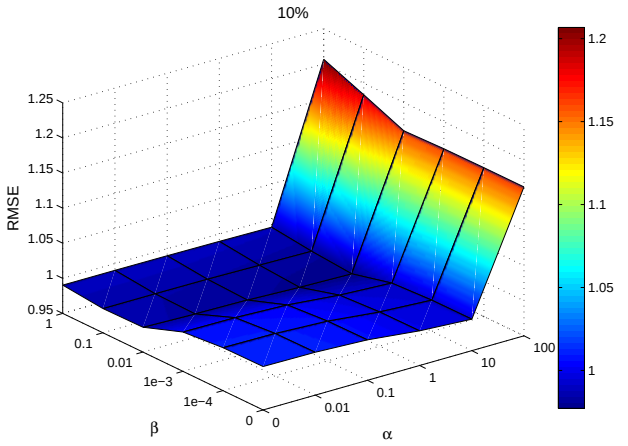
	Metrics	10%	50%	90%
Positive	*TopicMean*	0.5834	0.6738	0.6946
	UCF	0.6490	0.6922	0.7073
	MF	0.6911	0.6822	0.6888
	ScMF	0.6867	0.6932	0.7024
	SfMF	0.6918	0.6900	0.6991
	TcMF	0.7265	0.7034	0.7080
	ToMF	0.7232	0.7019	0.7066
	ScTcMF	0.7228	0.7042	0.7105
Negative	*TopicMean*	0.4808	0.4276	0.4049
	UCF	0.4553	0.4193	0.4013
	MF	0.3911	0.4735	0.4796
	ScMF	0.4302	0.4983	0.5067
	SfMF	0.4179	0.4932	0.4975
	TcMF	0.2672	0.4418	0.4702
	ToMF	0.2904	0.4427	0.4720
	ScTcMF	0.3040	0.4804	0.4965
Average	*TopicMean*	0.5321	0.5507	0.5498
	UCF	0.5522	0.5558	0.5543
	MF	0.5411	0.5779	0.5829
	ScMF	**0.5585**	**0.5958**	**0.6046**
	SfMF	0.5549	0.5916	0.5983
	TcMF	0.4969	0.5726	0.5891
	ToMF	0.5068	0.5722	0.5893
	ScTcMF	0.5134	0.5923	0.6035

which indicates that it is more difficult to predict negative opinions correctly. The matrix factorization models incorporating regularization constraints clearly outperform the state-of-the-art methods in no matter positive opinion prediction or negative opinion prediction. ScTcMF is the most precise one and also gains the comparable best results measured by recall and F1 score. In the situation that the training data size is extremely less than the test data size (e.g. only 10% training data), ScTcMF tends to predict more samples in the test set as positive, bringing about high recall and F1 score in positive opinion prediction but low recall and F1 score in negative opinion prediction.

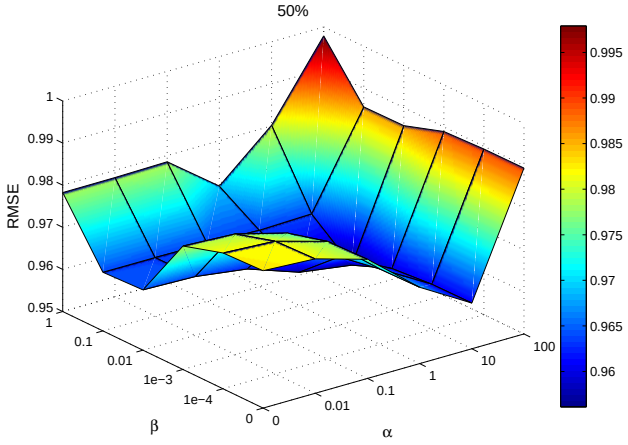
In spite of some limitations in this work, our proposed framework stably makes improvements with various evaluation metrics. Considering the inherent difficulty of our task, ScTcMF works not bad even though the prediction accuracy does not look so impressive.

6.3.5 Parameter Analysis

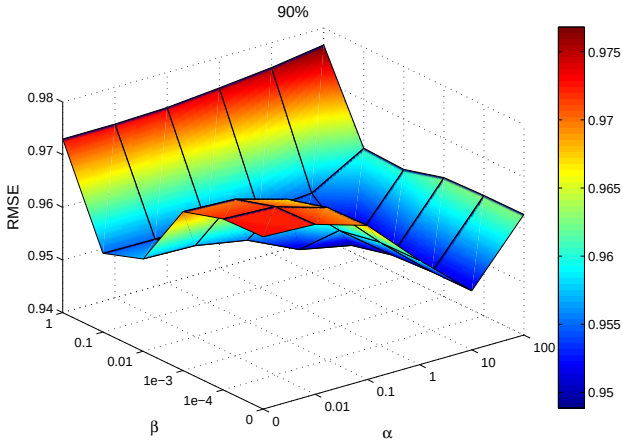
In the task of user interest prediction, we have analyzed the effects of social and topical context regularization on improving the performance of low-rank matrix factorization. In the task of user opinion prediction, we mainly investigate the impact of the social context regularization parameter α and the topical context regularization parameter β. As mentioned in Section 3, the regularization parameters are set to balance the reconstruction error in the original matrix factorization terms and in the regularization terms. Thus parameters α and β play important roles in controlling how much contribution the ScTcMF framework could gain from the regularization constraints of the social and topical context. Here we set the value of α to $\{0, 0.01, 0.1, 1, 10, 100\}$ to learn its impact on the prediction performance. In our case, β doesn't equal to α when the ScTcMF method achieves its best performance, so we vary β as $\{0, 1e-4, 1e-3, 0.01, 0.1, 1\}$ to show the impact of β.



(a)

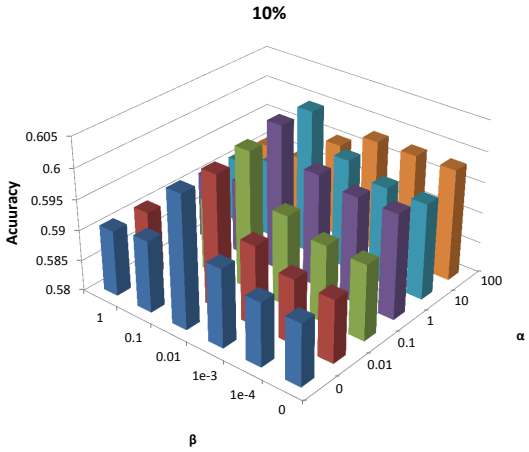


(b)

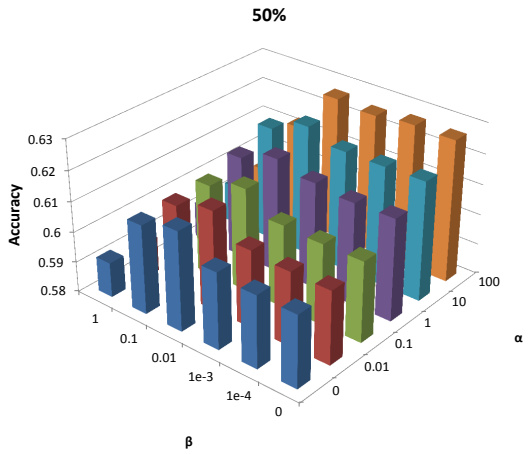


(c)

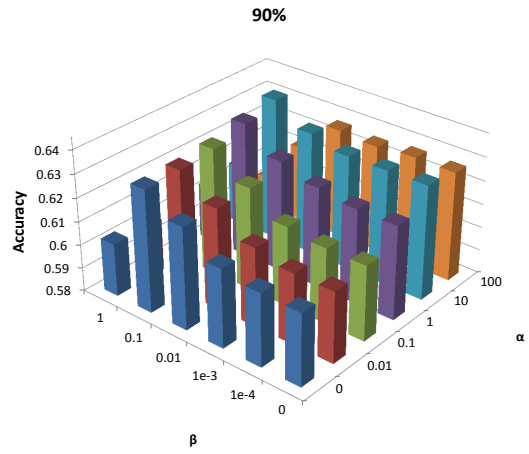
Figure 6.3: Impact of parameters α and β on RMSE.



(a)



(b)



(c)

Figure 6.4: Impact of parameters α and β on Accuracy.

The experimental results are shown in Figure 6.3 and Figure 6.4, which reveal the impact of α and β on RMSE and Accuracy respectively. To save space, we only show the results of the experiments using 10%, 50% and 90% training data, which are enough to help understand the trend of the impact of the regularization parameters with different sizes of training sets. Note that in this experiments, when $\alpha = \beta = 0$, it degrades to the MF model; when $\alpha > 0, \beta = 0$, it is actually the ScMF model; when $\alpha = 0, \beta > 0$, it becomes the TcMF model.

- The impact of α shares the similar trend as the impact of β. With too small parameter values, the impact of regularization constraints can be ignored, so the performance of the ScTcMF framework is almost the same as the basic MF model; with too large parameter values, the regularization terms will dominate the whole objective function and result in even worse performance. Only the appropriate α and β can lead to significant improvements.

- With the same β, as α varies from 0.01 to 10, the RMSE value decreases, and the Accuracy value increases in the meantime. We can see that the proposed framework achieves its best performance when $\alpha = 10$. When the value of α gets larger than 10, the performance becomes worse dramatically.

- Similarly, when we fix α, and vary β from small to large, the prediction performance of the ScTcMF framework becomes better at first, and then achieves the best, and becomes worse later. Using 50% and 90% of training data, the smallest RMSE values appear when $\beta = 0.1$, and the comparable small values appear when $\beta = 0.01$. However, the best results of Accuracy are obtained when $\beta = 0.01$. Comprehensively, we adopt $\beta = 0.01$ as the best regularization parameter of topical context in our experiments. Note that with $Train_{10\%}$, the impact of β on RMSE is not significant, and the results of Accuracy when using large β are even smaller than the results of Accuracy when $\beta = 0$. Perhaps the reason is that the correlations among topics are difficult to be explored

when the observed data is too sparse, so the regularization constraint of topical context cannot play its due role.

- In the task of user opinion prediction, the ScTcMF framework gets the best performance when $\alpha = 10, \beta = 0.01$.

Chapter 7

Conclusion and Future Work

In this dissertation, we focus on exploiting social and topical context for predicting user preference in microblogging. The main contributions of our work include: 1) We propose a general framework for incorporating social context and topical context as regularization constraints to help improve the performance of two user preference prediction tasks. 2) For predicting user interest, we exploit the characteristics of topical opinion distribution to describe topical context information, and further capture the weights between social friends under different opinion distribution topic patterns as social context information. 3) For predicting user opinion, we utilize content-based correlations among topics as topical context information, and social friend relationships between users as social context information. 4) The proposed ScTcMF framework is empirically evaluated on a real-world Twitter dataset, and the experimental results demonstrate that social and topical context can lead to improvements in both user interest prediction and user opinion prediction. We conclude the two user preference prediction tasks and present the future work in the following sections.

7.1 Conclusion

In the task of user interest prediction, we first propose the characteristics of topical opinion distribution, and give a vector representation for each topic. Then the similarities between opinion distribution vectors of topic pairs are calculated to describe topical context. We further divide topics into different patterns based on the three most important characteristics of topical opinion distribution to learn user interests in different topics, and extract the user information under different opinion distribution topic patterns to construct social context. Using the proposed ScTcMF framework, the experimental results on the collected real-world Twitter dataset exhibit its good performance.

In the task of user opinion prediction, social context is formulated according to the homophily theory. The similarities of users' opinions towards topics are measured for imposing the regularization constraint of social context. For incorporating topical context, we investigate the content-based correlations among the topics, and validate corresponding hypothesis to develop regularization constraint. Finally, experiments are carried out to evaluate the proposed ScTcMF framework, and the results show that ScTcMF framework performs better than the baseline methods.

7.2 Future Work

Some limitations we encountered in this work suggest directions for future work.

In this dissertation, we denote hashtags as the topics created by users. Although it has been adopted in previous topic-focused work, this topic detection approach actually has several problems. Hashtags in microblogging are labeled by users. On one hand, it means these hashtags can represent the targets that users pointed out in their posts. On the other hand, there is no unified standard for labeling hashtag, so different users may create different hashtags for the same topic, which will produce redundant topics and impact on formulating topical context for the

prediction tasks. Besides, abbreviation and the characteristic of no space also raise difficulties for accurate topic detection. Therefore, an effective and efficient approach to detect topics from the user-generated data should be proposed in the future.

We consider social friend relationships in microblogging to model social context in our current work. The prior similarities between user pairs are calculated to weight their relationships. In the future, more fine-grained features such as if two people live close to each other, if they are active in the same time, could also be considered to weight the relationships between microblogging users.

In user interest prediction, we explore the interest similarities between users under different topic patterns, which are helpful to understand the implicit relationships between them in more detail. The findings of the detailed user relationship mining from different patterns could be applied to other future tasks like friend recommendation in microblogging.

In user interest prediction, we observe from the experimental results that the RMSE values of the proposed framework significantly decrease comparing with the baseline methods. However, the improvement of accuracy is not so dramatic. The possible reason is that the two states of positive and negative is not enough to describe user opinion. To predict multiple states of user opinion, is a challenging issue worth learning in the future.

Bibliography

[1] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)*, 23(1):103–145, 2005.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[3] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[4] S. Asur and B. A. Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[5] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.

[6] L. Barbosa and J. Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[7] J. Benhardus and J. Kalita. Streaming trend detection in twitter. *International Journal of Web Based Communities*, 9(1):122–139, 2013.

[8] D. Blei and J. Lafferty. Correlated topic models. *Advances in neural information processing systems*, 18:147, 2006.

[9] J. Bollen, B. Gonçalves, G. Ruan, and H. Mao. Happiness is assortative in online social networks. *Artificial Life*, 17(3):237–251, 2011.

[10] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[11] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 450–453, 2011.

[12] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 43–52. Morgan Kaufmann Publishers Inc., 1998.

[13] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1185–1194. ACM, 2010.

[14] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 661–670. ACM, 2012.

[15] M. Chu, F. Diele, R. Plemmons, and S. Ragni. Optimality, computation, and interpretation of nonnegative matrix factorizations. In *SIAM Journal on Matrix Analysis*. Citeseer, 2004.

[16] A. S. Das, M. Datar, A. Garg, and S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, pages 271–280. ACM, 2007.

[17] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics, 2010.

[18] J. H. Davis, T. Kameda, C. Parks, M. Stasson, and S. Zimmerman. Some social mechanics of group decision making: The distribution of opinion, polling sequence, and implications for consensus. *Journal of Personality and Social Psychology*, 57(6):1000, 1989.

[19] E. Diaz-Aviles, L. Drumond, Z. Gantner, L. Schmidt-Thieme, and W. Nejdl. What is happening right now... that interests me? In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012.

[20] C. Ding, T. Li, and M. I. Jordan. Nonnegative matrix factorization for combinatorial optimization: Spectral clustering, graph matching, and clique finding. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 183–192. IEEE, 2008.

[21] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 617–624. ACM, 2005.

[22] H. Gao, J. Tang, X. Hu, and H. Liu. Exploring temporal effects for location recommendation on location-based social networks. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 93–100. ACM, 2013.

[23] M. A. Ghazanfar, A. Prügel-Bennett, and S. Szedmak. Kernel-mapping recommender system algorithms. *Information Sciences*, 208:81–104, 2012.

[24] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical report, 2009.

[25] S. A. Golder and M. W. Macy. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science*, 333(6051):1878–1881, 2011.

[26] Q. Gu, J. Zhou, and C. Ding. Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs. In *SIAM SDM*, pages 199–210, 2010.

[27] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 199–206. ACM, 2010.

[28] J. Hannon, K. McCarthy, and B. Smyth. Finding useful users on twitter: twittomender the followee recommender. In *Advances in Information Retrieval*, pages 784–787. Springer, 2011.

[29] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 230–237. ACM, 1999.

[30] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2000.

[31] T. Hofmann. Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 259–266. ACM, 2003.

[32] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115, 2004.

[33] T. Hofmann and D. Hartmann. Collaborative filtering with privacy via factor analysis. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, pages 791–795, 2005.

[34] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.

[35] X. Hu, J. Tang, H. Gao, and H. Liu. Actnet: Active learning for networked texts in microblogging. In *SDM*, pages 306–314. SIAM, 2013.

[36] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[37] L. Huo-yao and L. Gong-shen. Prediction on semantic orientation of texts based on topic correlation [j]. *Information Security and Communications Privacy*, 3:033, 2009.

[38] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[39] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[40] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 151–160, 2011.

[41] R. Jin, J. Y. Chai, and L. Si. An automatic weighting scheme for collaborative filtering. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 337–344. ACM, 2004.

[42] H. Kanayama and T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363. Association for Computational Linguistics, 2006.

[43] P. Kapanipathi, F. Orl, and A. Sheth. Personalized filtering of the twitter stream. In *SPIM, volume 781 of CEUR Workshop Proceedings, 6C13. CEUR-WS.org*, 2011.

[44] S.-M. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[45] Y. Kim and K. Shim. Twitobi: A recommendation system for twitter using probabilistic modeling. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 340–349. IEEE, 2011.

[46] I. King, J. Li, and K. Chan. A brief survey of computational approaches in social computing. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 1625–1632. IEEE, 2009.

[47] A. Kohrs and B. Merialdo. Clustering for collaborative filtering applications. In *In Computational Intelligence for Modelling, Control & Automation. IOS*. Citeseer, 1999.

[48] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[49] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM, 2008.

[50] E. Kouloumpis, T. Wilson, and J. Moore. Twitter sentiment analysis: The good the bad and the omg. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[51] O. Kucuktunc, B. B. Cambazoglu, I. Weber, and H. Ferhatosmanoglu. A large-scale sentiment analysis for yahoo! answers. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 633–642. ACM, 2012.

[52] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[53] D. Lee, H. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[54] X. Li and T. Murata. A hybrid method using multidimensional clustering-based collaborative filtering to improve recommendation diversity. *IEEJ Transactions on Electronics, Information and Systems*, 133(4):749–755, 2013.

[55] C. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[56] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *Proceeding of the 18th ACM conference on Information and knowledge management*, pages 375–384. ACM, 2009.

[57] J. Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.

[58] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 422–429. ACM, 2011.

[59] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[60] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700. ACM, 2010.

[61] H. Ma, I. King, and M. R. Lyu. Effective missing data prediction for collaborative filtering. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 39–46. ACM, 2007.

[62] H. Ma, H. Yang, M. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 931–940. ACM, 2008.

[63] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 287–296. ACM, 2011.

[64] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[65] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. In *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 492–508. Springer, 2004.

[66] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[67] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, pages 171–180. ACM, 2007.

[68] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, pages 73–80. ACM, 2010.

[69] B. N. Miller, J. A. Konstan, and J. Riedl. Pocketlens: Toward a personal recommender system. *ACM Transactions on Information Systems (TOIS)*, 22(3):437–476, 2004.

[70] B. OConnor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129, 2010.

[71] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC*, volume 2010, 2010.

[72] R. Pan, Y. Zhou, B. Cao, N. N. Liu, R. Lukose, M. Scholz, and Q. Yang. One-class collaborative filtering. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 502–511. IEEE, 2008.

[73] Y. Pan, F. Cong, K. Chen, and Y. Yu. Diffusion-aware personalized social update recommendation. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 69–76. ACM, 2013.

[74] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[75] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[76] B. Pang and L. Lee. Sentiment analysis and opinion mining. *Foundations and Trends in Information Retrieval*, (1):1–135, 2008.

[77] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[78] M. Papagelis, D. Plexousakis, and T. Kutsuras. Alleviating the sparsity problem of collaborative filtering using trust inferences. In *Trust management*, pages 224–239. Springer, 2005.

[79] J. Park, M. Cha, H. Kim, and J. Jeong. Managing bad news in social media: A case study on dominos pizza crisis. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[80] R. Pfitzner, A. Garas, and F. Schweitzer. Emotional divergence influences information spreading in twitter. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[81] S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE, 2010.

[82] S. Puntheeranurak and H. Tsuji. An improved hybrid recommender system using multi-based clustering method. *IEEJ Transactions on Electronics, Information and Systems*, 129:125–132, 2009.

[83] F. Ren and Y. Wu. Predicting user-topic opinions in twitter with social and topical context. *IEEE Transactions on Affective Computing*, 4(4):412–424, 2013.

[84] J. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[85] F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.

[86] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20:1257–1264, 2008.

[87] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM, 2001.

[88] G. Shani and A. Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[89] M. Skoric, N. Poor, P. Achananuparp, E.-P. Lim, and J. Jiang. Tweets and votes: A study of the 2011 singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2583–2591. IEEE, 2012.

[90] S. S. Sohail, J. Siddiqui, and R. Ali. Book recommendation system using opinion mining technique. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1609–1614. IEEE, 2013.

[91] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, pages 53–63. Association for Computational Linguistics, 2011.

[92] N. Srebro, T. Jaakkola, et al. Weighted low-rank approximations. In *ICML*, volume 3, pages 720–727, 2003.

[93] A. Stavrianou and C. Brun. Expert recommendations based on opinion mining of user-generated product reviews. *Computational Intelligence*, 2013.

[94] C. Strapparava and A. Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[95] J.-H. Su, H.-H. Yeh, P. S. Yu, and V. S. Tseng. Music recommendation using content and context information mining. *Intelligent Systems, IEEE*, 25(1):16–26, 2010.

[96] X. Su and T. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4, 2009.

[97] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1397–1405. ACM, 2011.

[98] J. Tang, H. Gao, X. Hu, and H. Liu. Exploiting homophily effect for trust prediction. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[99] J. Tang, H. Gao, and H. Liu. mtrust: Discerning multi-faceted trust in a connected world. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 93–102. ACM, 2012.

[100] J. Tang, Y. Zhang, J. Sun, J. Rao, W. Yu, Y. Chen, and A. Fong. Quantitative study of individual emotional states in social networks. *Affective Computing, IEEE Transactions on*, 3(2):132–144, 2012.

[101] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[102] M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63:163–173, 2012.

[103] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[104] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international aaai conference on weblogs and social media*, pages 178–185, 2010.

[105] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics, 2002.

[106] L. H. Ungar and D. P. Foster. Clustering methods for collaborative filtering. In *AAAI Workshop on Recommendation Systems*, volume 1, 1998.

[107] F. Wang and L. Chen. Recommendation based on mining product reviewers preference similarity network. In *Proceedings of 6th SNAKDD workshop (Page: 166 Year of Publication: 2012 ISBN: 78-1-4503-1544-9)*, 2012.

[108] Y.-X. Wang and H. Xu. Stability of matrix factorization for collaborative filtering. *arXiv preprint arXiv:1206.4640*, 2012.

[109] R. W. White and J. M. Jose. A study of topic similarity measures. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 520–521. ACM, 2004.

[110] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

[111] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics, 2005.

[112] G.-R. Xue, C. Lin, Q. Yang, W. Xi, H.-J. Zeng, Y. Yu, and Z. Chen. Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 114–121. ACM, 2005.

[113] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, and J.-M. Ho. Author name disambiguation for citations using topic and web correlation. *Research and Advanced Technology for Digital Libraries*, pages 185–196, 2008.

[114] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pages 427–434. IEEE, 2003.

[115] K. Yu, S. Zhu, J. Lafferty, and Y. Gong. Fast nonparametric matrix factorization for large-scale collaborative filtering. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 211–218. ACM, 2009.

[116] R. Zafarani, W. D. Cole, and H. Liu. Sentiment propagation in social networks: A case study in livejournal. In *Advances in Social Computing*, pages 413–420. Springer, 2010.

[117] J. Zhu, H. Wang, M. Zhu, B. K. Tsou, and M. Ma. Aspect-based opinion polling from customer reviews. *Affective Computing, IEEE Transactions on*, 2(1):37–49, 2011.

[118] J. Zhu, C. Zhang, and M. Ma. Multi-aspect rating inference with aspect-based segmentation. *IEEE Transactions on Affective Computing*, 3:469–481, 2012.